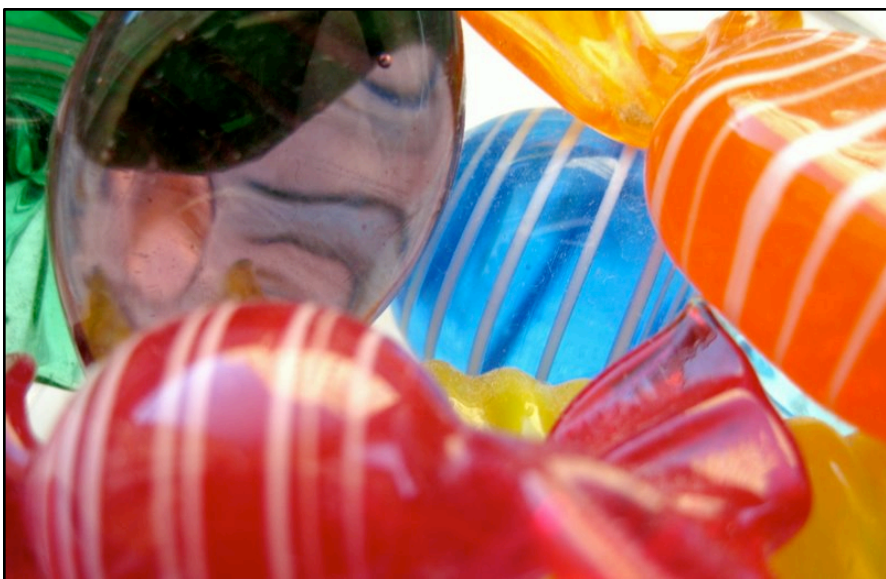




Understanding Protein Function on a Genome-scale through the Analysis of Molecular Networks

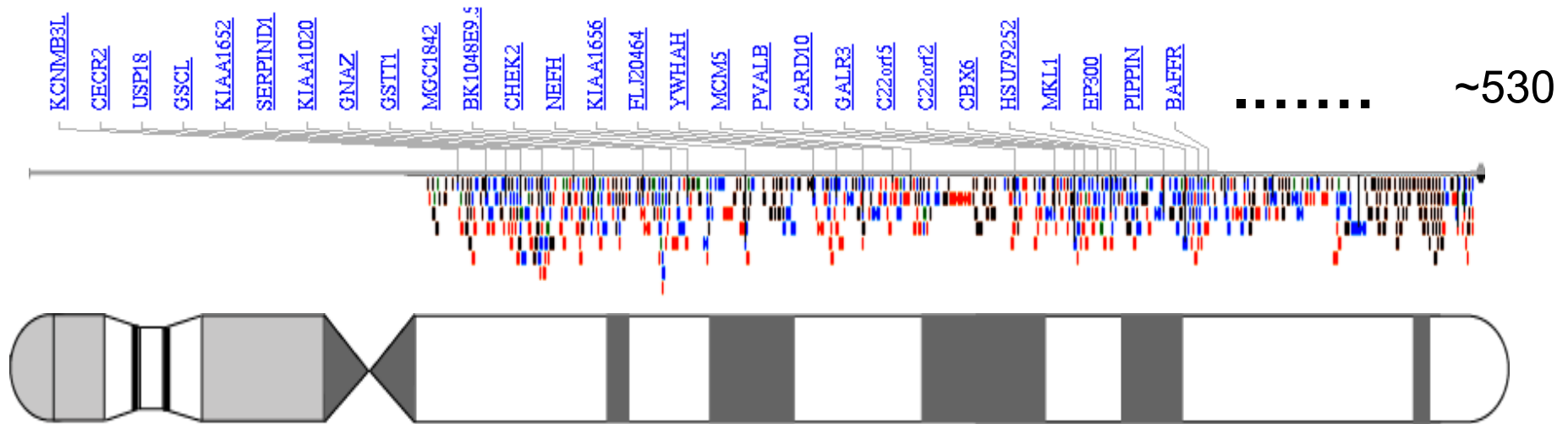
Mark B Gerstein
Yale



slides at
Lectures.GersteinLab.org

(See Last Slide for References
& More Info.)

The problem: Grappling with Function on a Genome Scale?



- 250 of ~530
originally characterized on chr. 22
[Dunham et al. Nature (1999)]
- >25K Proteins in Entire Human Genome
(with alt. splicing)

Traditional single molecule way to integrate evidence & describe function

EF2_YEAST

Descriptive Name:
Elongation Factor 2

Lots of references
to papers

Summary sentence describing function:
This protein promotes the GTP-dependent translocation of the nascent protein chain from the A-site to the P-site of the ribosome.

UniProt

the universal protein knowledgebase

Home About UniProt Getting Started Searches/Tools Databases Support/Documentation

Text Search UniProt Knowledgebase

General information about the UniProt/Swiss-Prot entry	
Entry name	EF2_YEAST
Primary accession number	P32324
Entered in Swiss-Prot	Release 27, 01-OCT-1993
Sequence was last modified	Release 27, 01-OCT-1993
Annotations were last modified	Release 47, 01-MAY-2005

Protein description	
Protein name	Elongation factor 2
Synonyms	EF-2

References	
[1]	NUCLEOTIDE SEQUENCE (EFT1 AND EFT2). MEDLINE=92112760; PubMed=1730643; [NCBI, ExPASy, EBI, Israel, Japan] Perentesis J.P., Phan L.D., Laporte D.C., Livingston D.M., Bodley J.W.; "Saccharomyces cerevisiae elongation factor 2. Genetic cloning, characterization of expression, and G-domain modeling."

Comments	
FUNCTION	This protein promotes the GTP-dependent translocation of the nascent protein chain from the A-site to the P-site of the ribosome.
SUBCELLULAR LOCATION	Cytoplasmic.

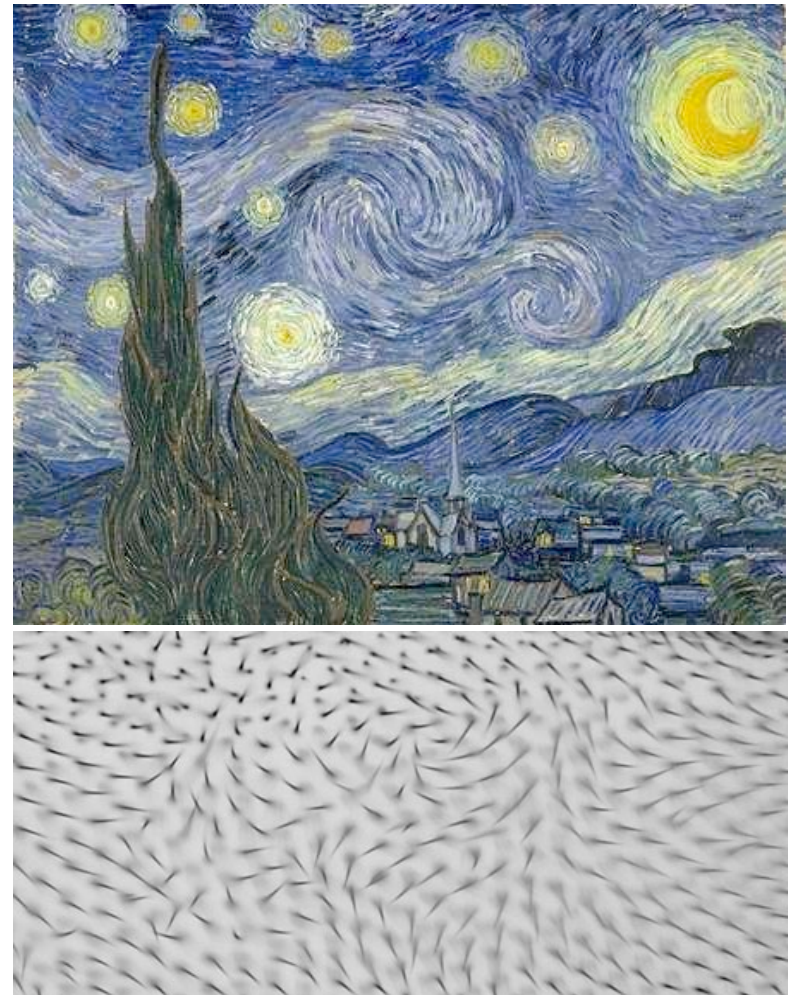
DIR Δ41778 Δ41778

Some obvious issues in scaling single molecule definition to a genomic scale

- Fundamental complexities
 - ◇ Often >2 proteins/function
 - ◇ Multi-functionality:
2 functions/protein
 - ◇ Role Conflation:
molecular, cellular, phenotypic

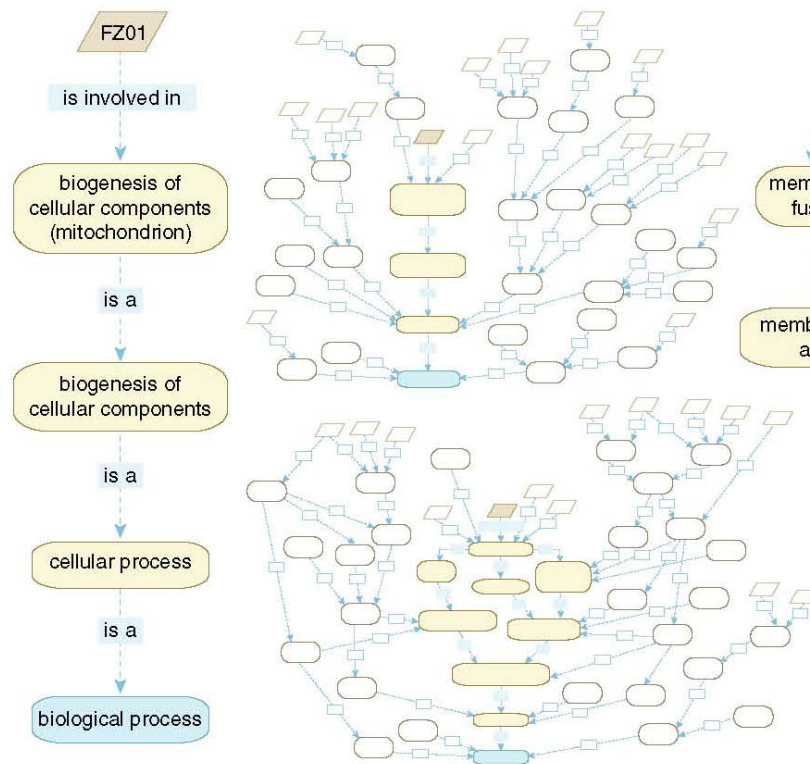
Some obvious issues in scaling single molecule definition to a genomic scale

- Fundamental complexities
 - ◇ Often >2 proteins/function
 - ◇ Multi-functionality:
2 functions/protein
 - ◇ Role Conflation:
molecular, cellular, phenotypic
- Fun terms... but do they scale?....
 - ◇ **Starry night** (P Adler, '94)

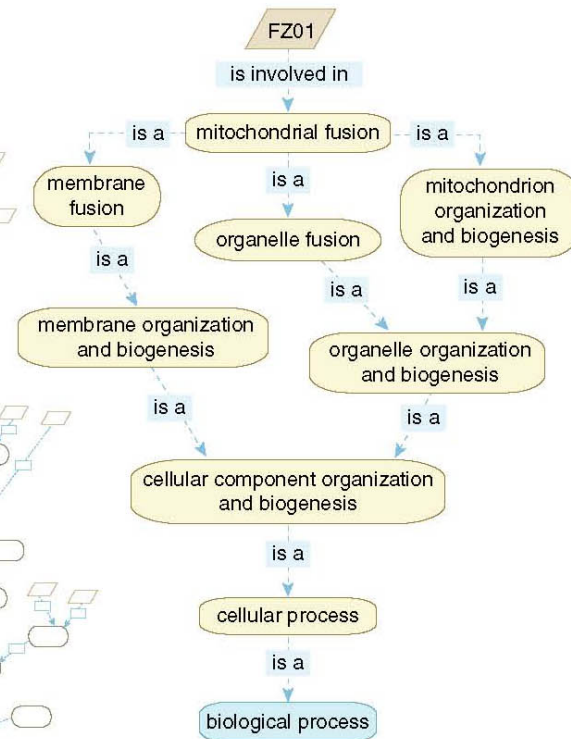


[Seringhaus et al. GenomeBiology (2008)]

Hierarchies & DAGs of controlled-vocab terms but still have issues...



MIPS (Mewes et al.)



GO (Ashburner et al.)

Towards Developing Standardized Descriptions of Function

- Subjecting each gene to standardized expt. and cataloging effect
 - ◊ KOs of each gene in a variety of std. conditions => phenotypes
 - ◊ Std. binding expts for each gene (e.g. prot. chip)

- Function as a vector

ector

nucleic acids

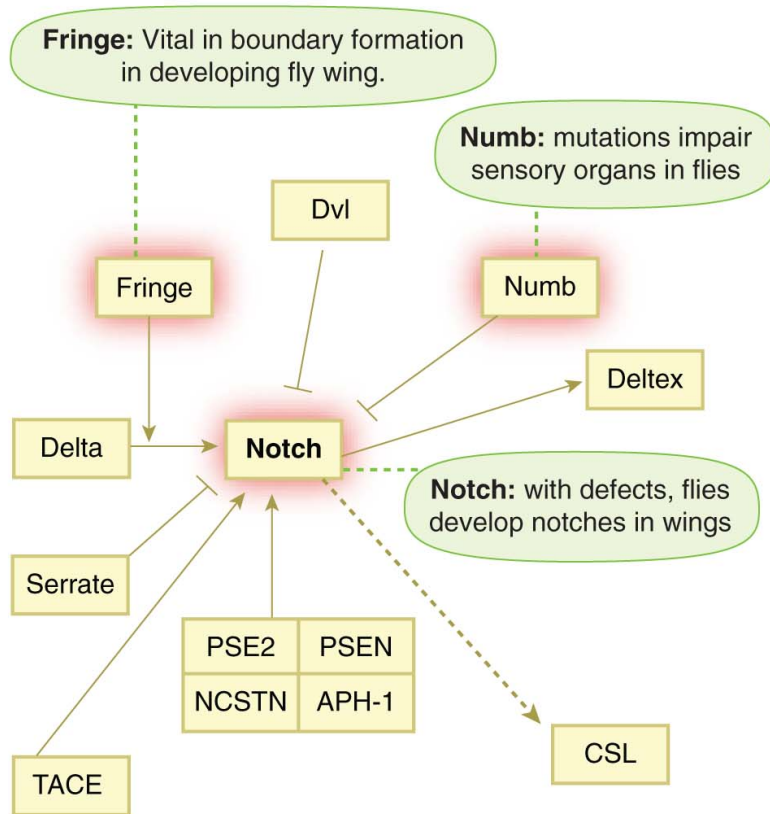
small molecules

proteins

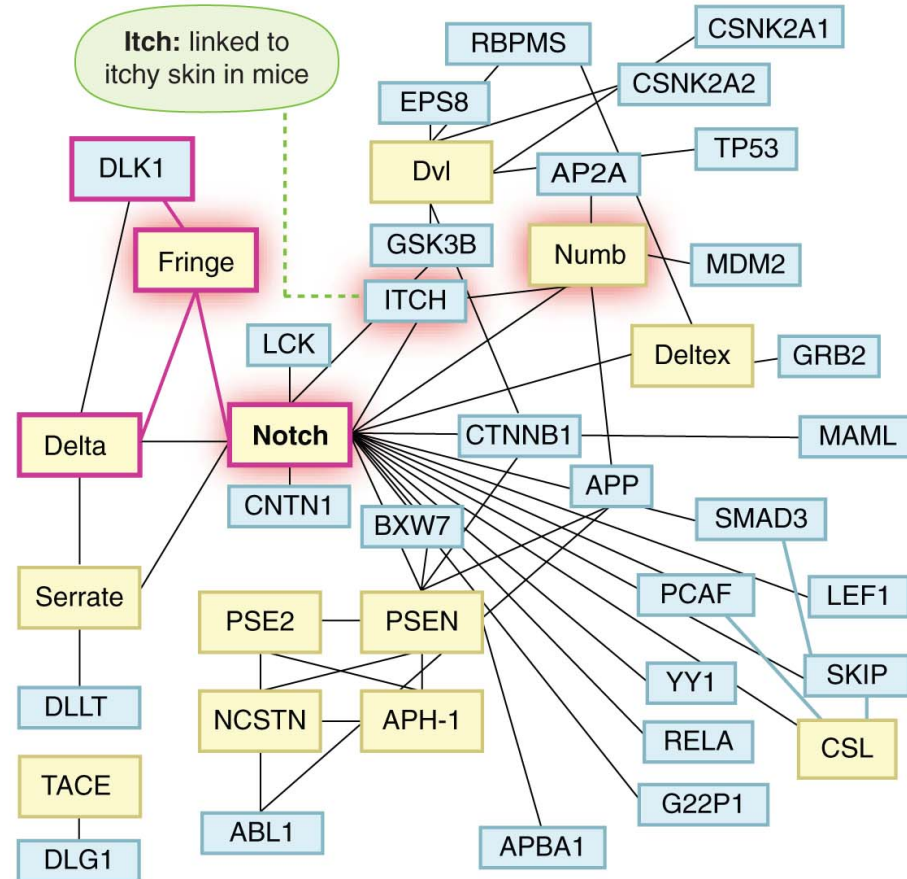
	DNA	RNA	ATP	Metal	CoA	NAD	G protein	CDC28	Calmodulin
protein 1	1.0	0	0	0	0	0	0	0	0
protein 2	0	0.9	0	0	0	0	0	0	0
protein 3	1.0	0	1.0	0	0	0	0	0	0
protein 4	0	0	0	0	0.8	0	0	0	1.0
protein 5	1.0	0	0	0	0	0	0	0.9	0
protein 6	0.9	0				
protein 7	0	0.8				
.....

Interaction Vectors [Lan et al, IEEE 90:1848]

Networks (Old & New)

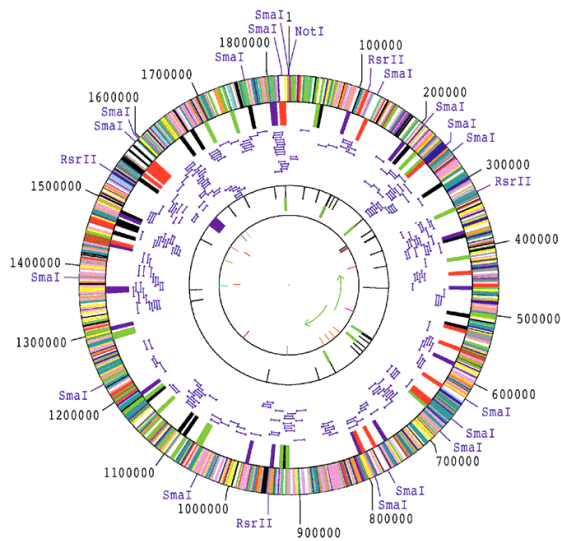


Classical KEGG pathway



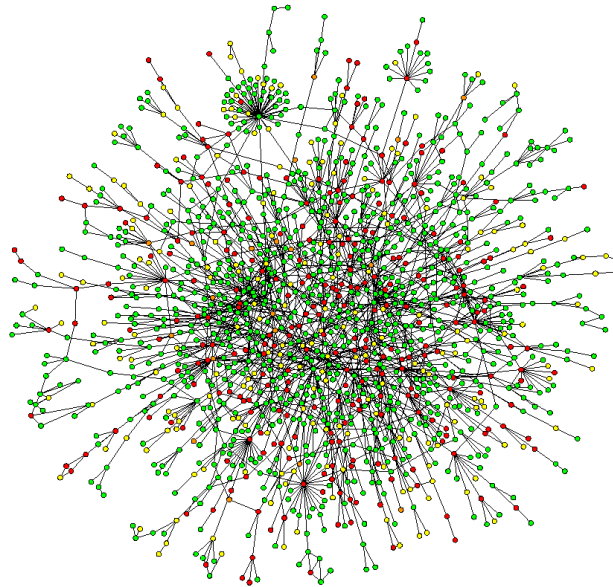
Same Genes in High-throughput Network

Networks occupy a midway point in terms of level of understanding



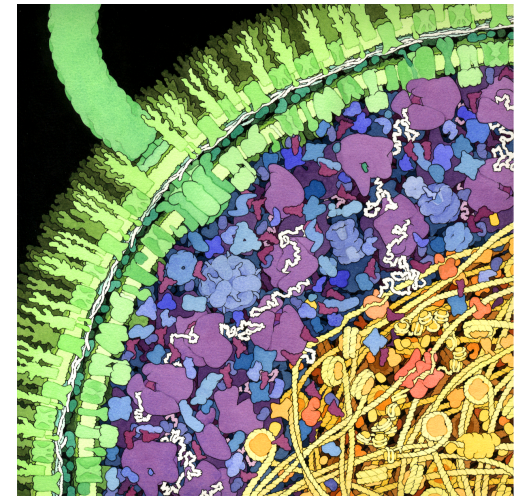
1D: Complete
Genetic Partslist

[Fleischmann et al., Science, 269 :496]



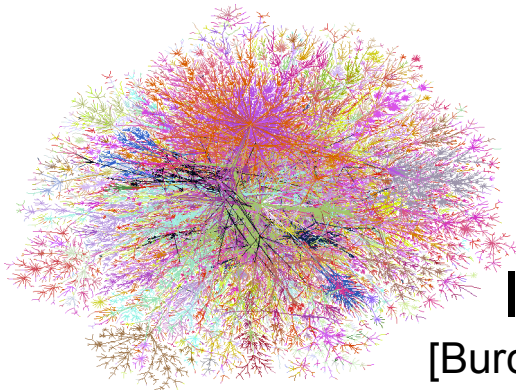
~2D: Bio-molecular
Network
Wiring Diagram

[Jeong et al. Nature, 41:411]

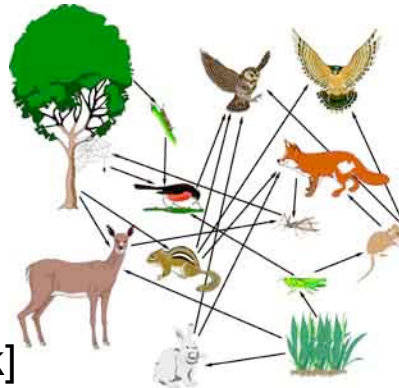


3D: Detailed
structural
understanding of
cellular machinery

Networks as a universal language



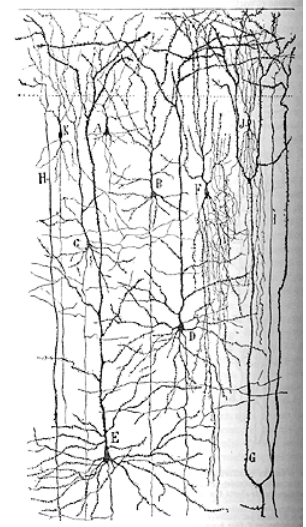
Internet
[Burch & Cheswick]



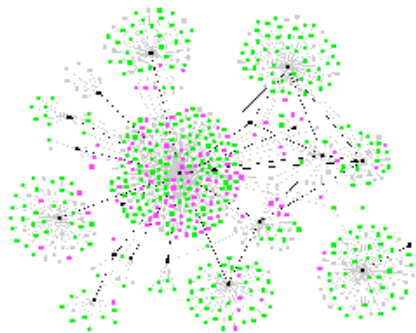
Food Web



Electronic
Circuit



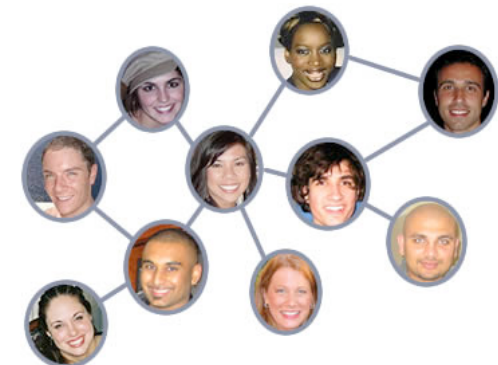
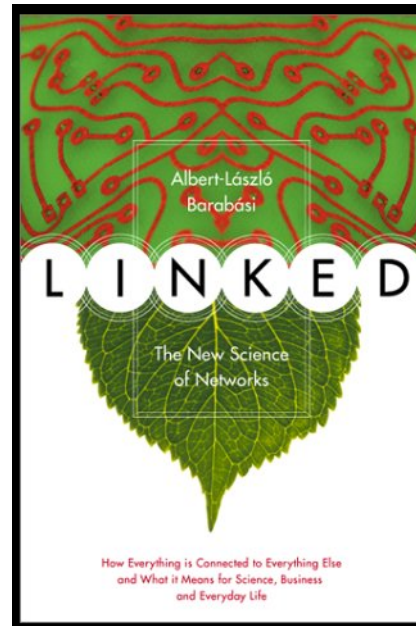
Neural Network
[Cajal]



Disease
Spread
[Krebs]



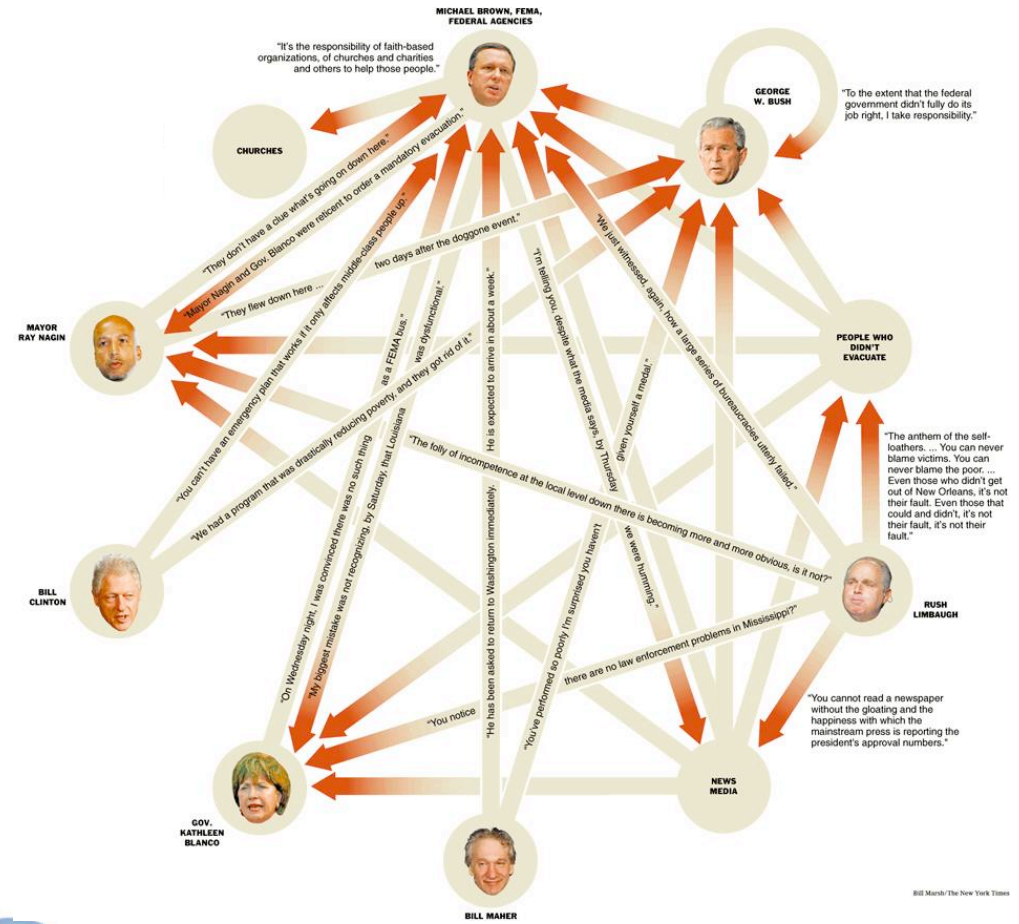
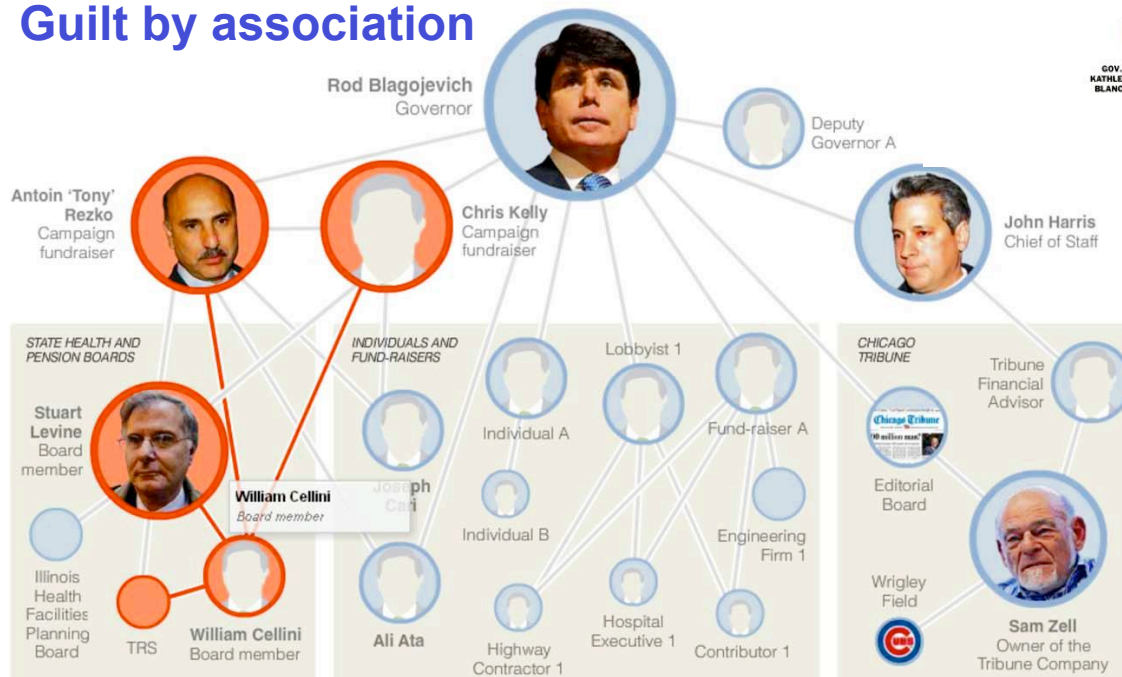
Protein
Interactions
[Barabasi]



Social Network

Using the position in networks to describe function

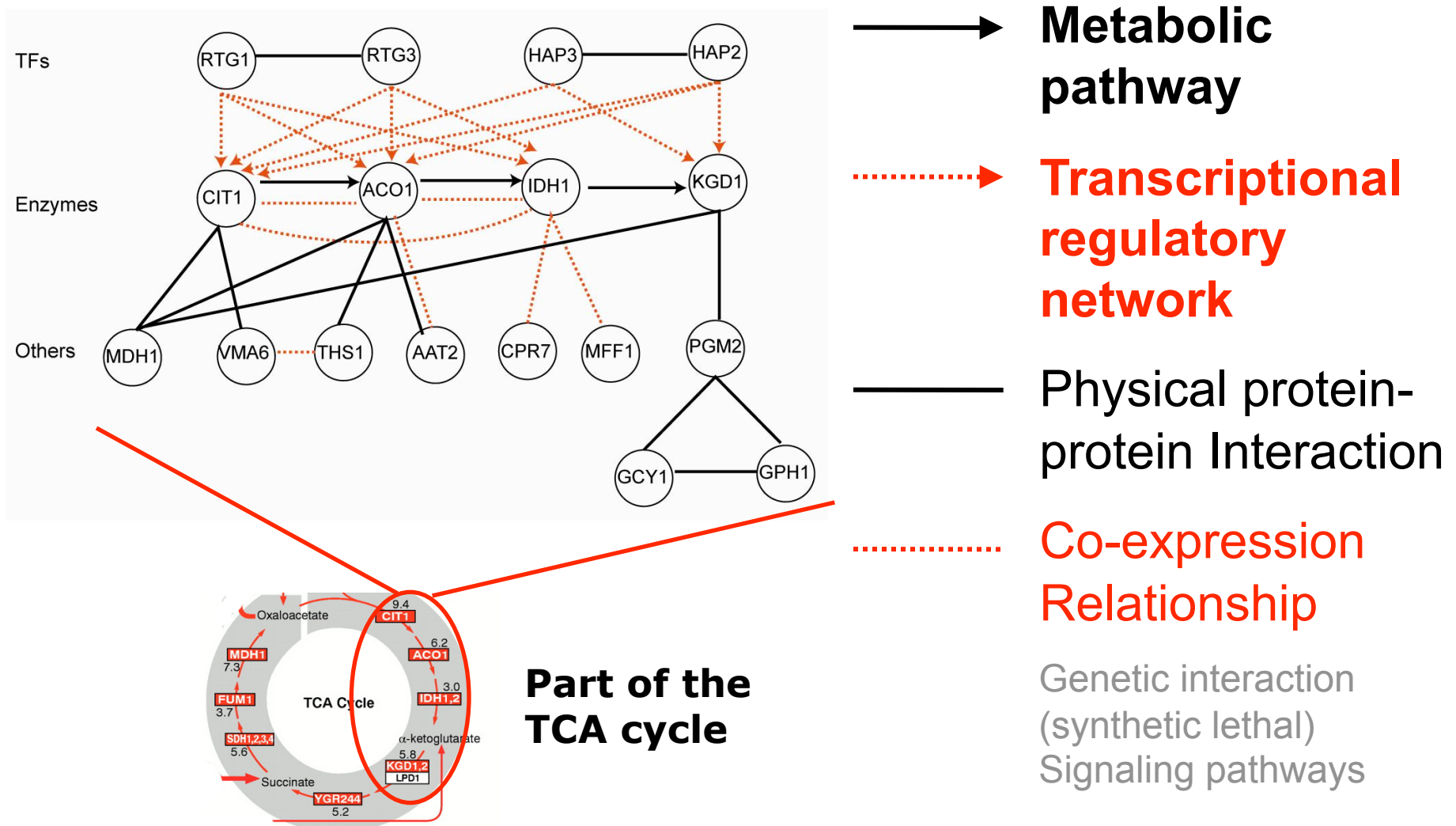
Guilt by association



Finding the causal regulator (the "Blame Game")

[NY Times, 2-Oct-05, 9-Dec-08]

Combining networks forms an ideal way of integrating diverse information



Outline: Molecular Networks

- Why Networks?
- Predicting Networks (yeast)
 - ◇ Propagating known information
- Dynamics & Variation of Networks
 - ◇ Across cellular states (yeast)
 - ◇ Across environments (in prokaryotes)
- Protein Networks & Variation (yeast & humans)



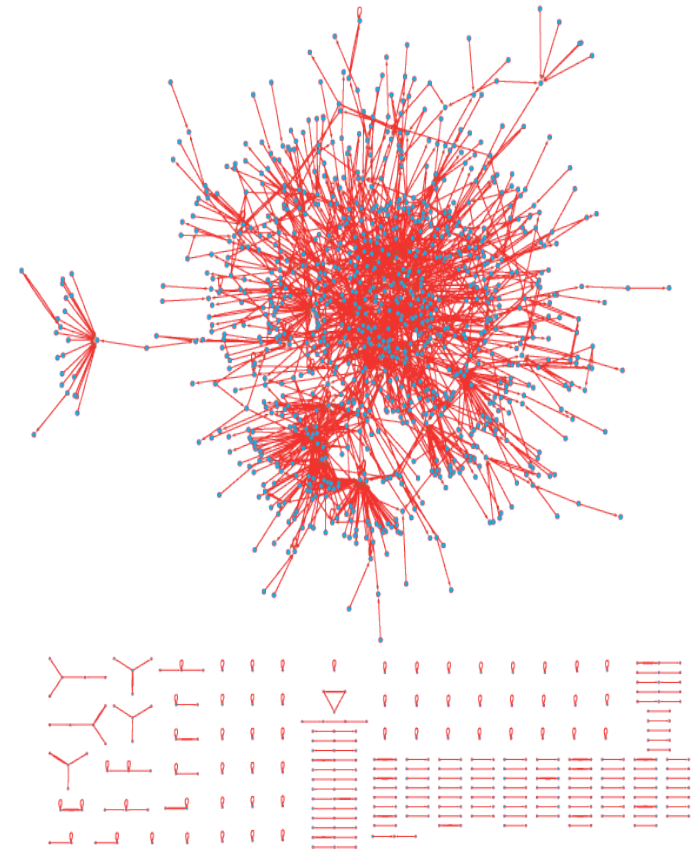
Example: yeast PPI network

Actual size:

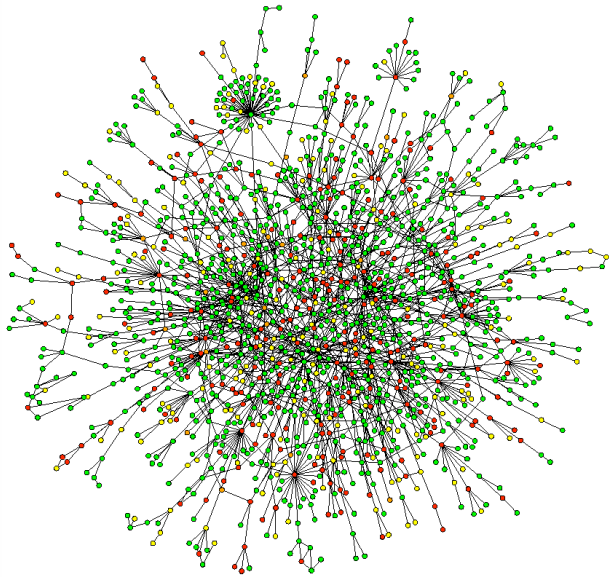
- ◇ ~6,000 nodes
→ Computational cost: ~18M pairs
- ◇ Estimated ~15,000 edges
→ Sparseness: 0.08% of all pairs
(Yu et al., 2008)

Known interactions:

- ◇ Small-scale experiments: accurate but few
→ Overfitting: ~5,000 in BioGRID, involving ~2,300 proteins
- ◇ Large-scale experiments: abundant but noisy
→ Noise: false +ve/-ve for yeast two-hybrid data up to 45% and 90% (Huang et al., 2007)



Types of Networks



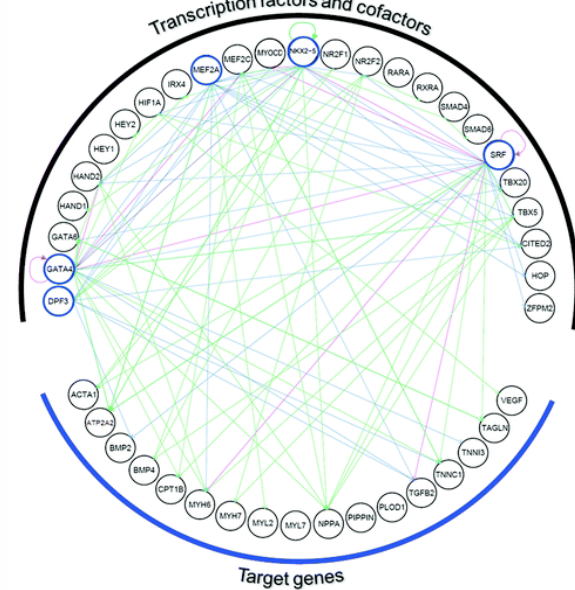
Interaction networks

Nodes: proteins or genes
Edges: interactions

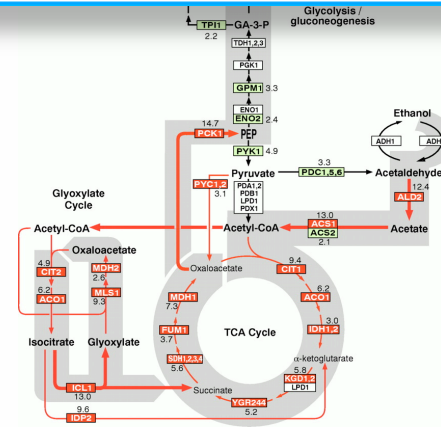
[Horak, et al, Genes & Development, 16:3017-3033]

[DeRisi, Iyer, and Brown, Science, 278:680-686]

[Jeong et al, Nature, 41:411]



Regulatory networks



Metabolic networks

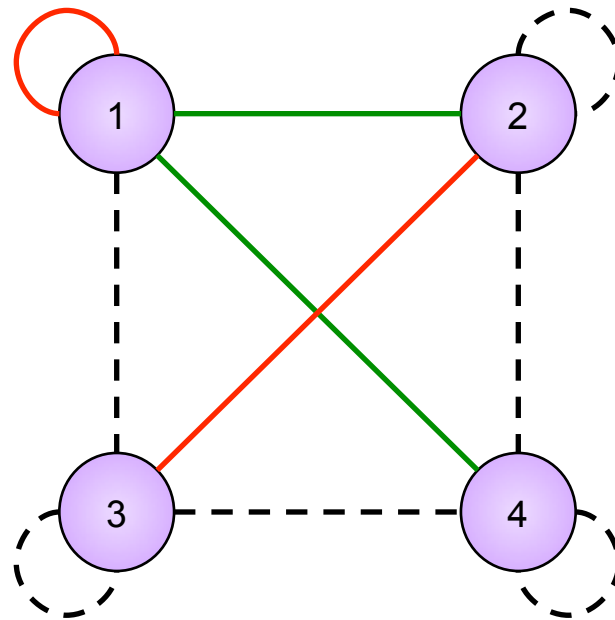
Predicting Networks

How do we construct large molecular networks?

From extrapolating correlations between functional genomics data with fairly small sets of known interactions, making best use of the known training data.



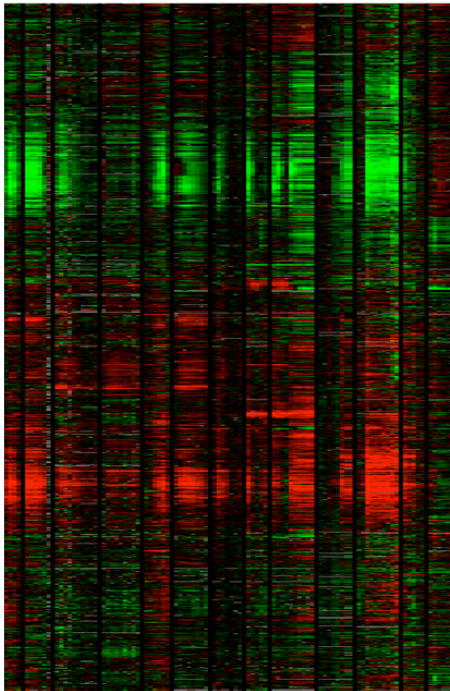
Network prediction: known information



- Known interactions
- Known non-interactions
- - - Unknown

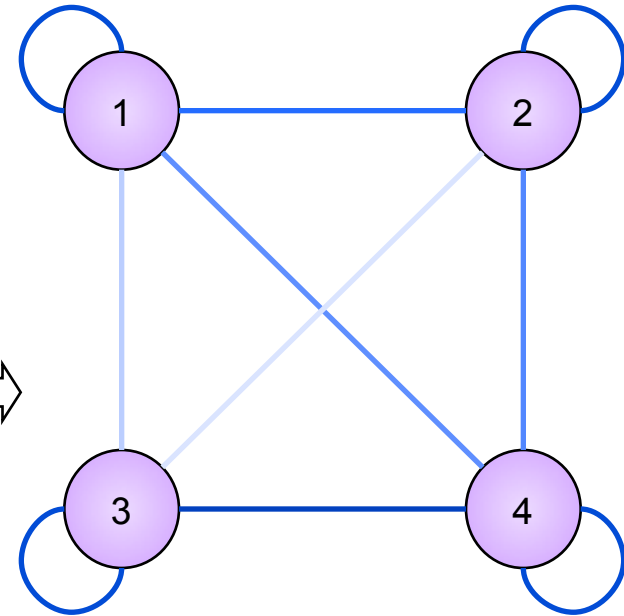
Network prediction: features

- Example 1: gene expression



Gasch et al., 2000

$$\begin{aligned} &\Rightarrow x_1 = (0.2, 2.4, 1.5, \dots) \\ &x_2 = (0.8, 2.2, 1.5, \dots) \\ &\Rightarrow x_3 = (4.3, 0.1, 7.5, \dots) \Rightarrow \\ &\dots \\ &\text{sim}(x_1, x_2) = 0.62 \\ &\text{sim}(x_1, x_3) = -0.58 \\ &\dots \end{aligned}$$

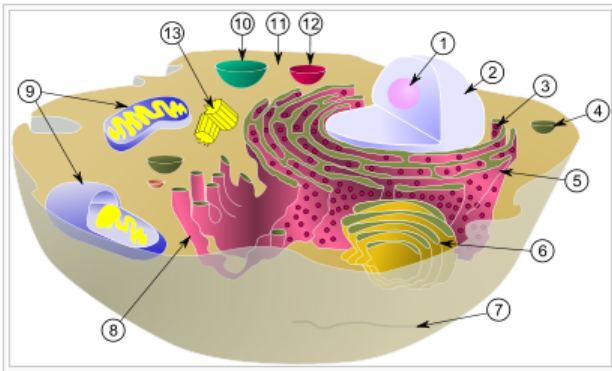


Similarity scale:



Network prediction: features

- Example 2: sub-cellular localization



<http://www.scq.ubc.ca/wp-content/yeasttwohybridtranscript.gif>

$x_1 = (1, 1, 0, 0, \dots)$

$x_2 = (1, 1, 1, 0, \dots)$

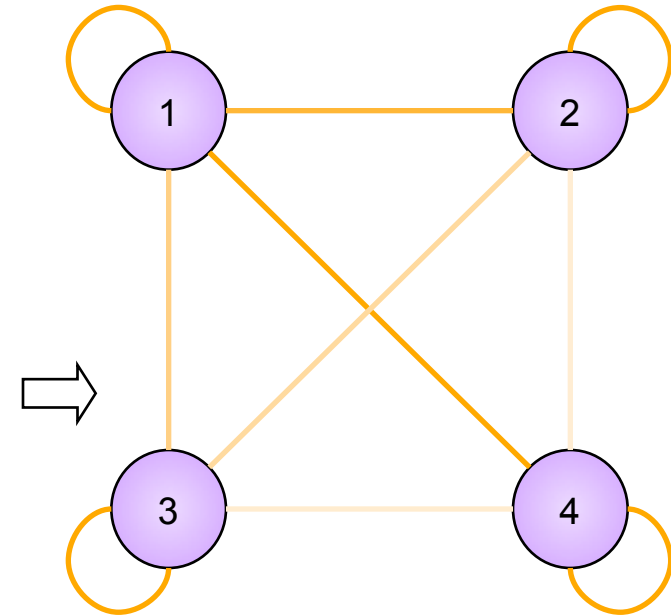
$x_3 = (1, 0, 1, 0, \dots)$

...

$\text{sim}(x_1, x_2) = 0.81$

$\text{sim}(x_1, x_3) = 0.12$

...



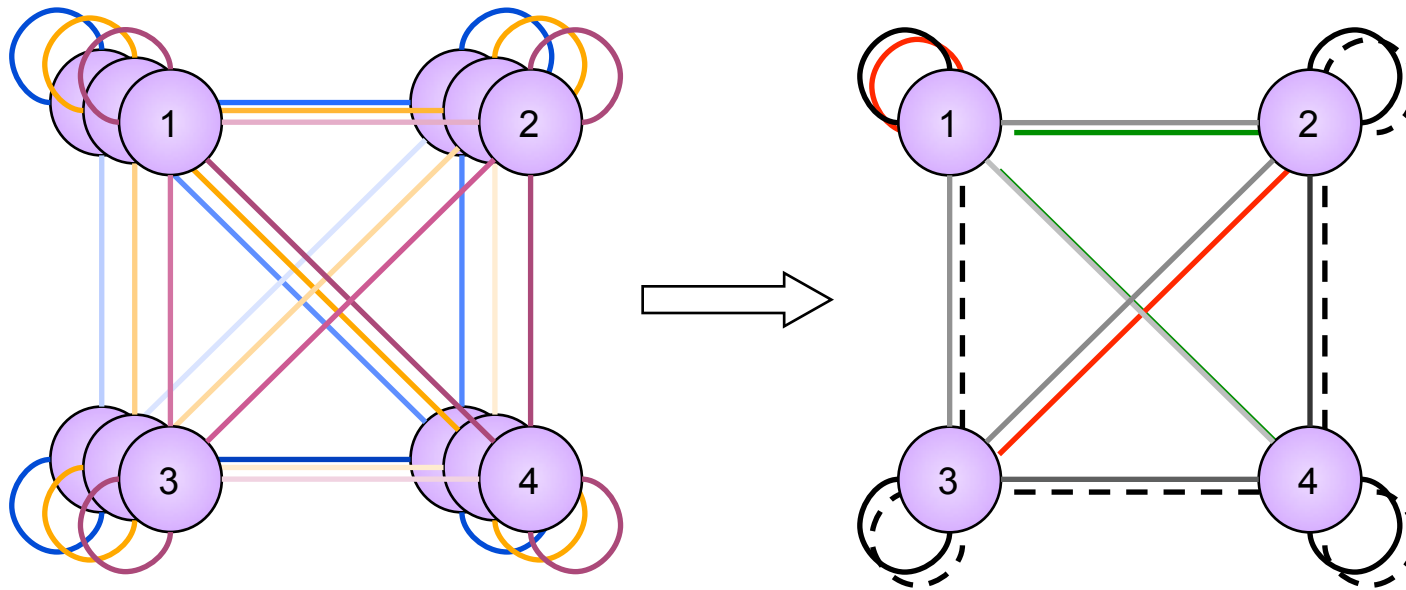
Similarity scale:

1



-1

Network prediction: data integration



Learning methods

An endless list:

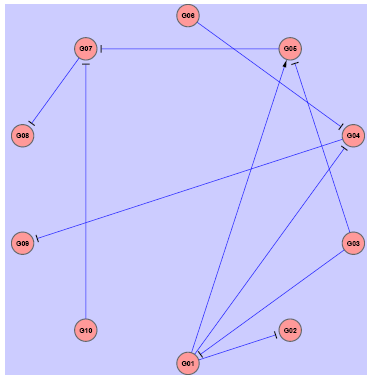
- Docking (e.g. Schoichet and Kuntz 1991)
- Evolutionary (e.g. Ramani and Marcotte, 2003)
- Topological (e.g. Yu et al., 2006)
- Bayesian (e.g. Jansen et al., 2003)
- **Kernel methods**
 - ◇ Global modeling:
 - em (Tsuda et al., 2003)
 - kCCA (Yamanishi et al., 2004)
 - kML (Vert and Yamanishi, 2005)
 - Pairwise kernel (Pkernel) (Ben-Hur and Noble, 2005)
 - ◇ Local modeling:
 - Local modeling (Bleakley et al., 2007)

Let's compare in a public challenge!

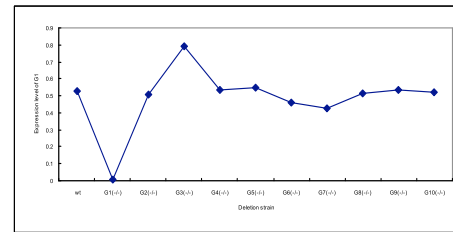
(DREAM: Dialogue for Reverse Engineering Assessment and Methods)

DREAM3: *in silico* regulatory network reconstruction

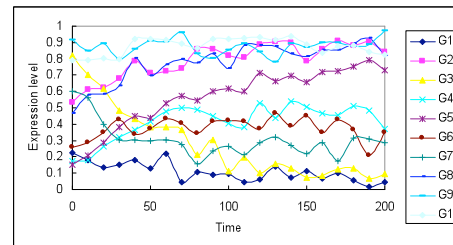
Actual network



Expression data

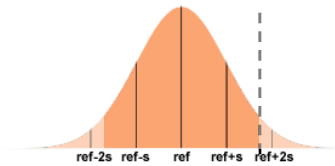


Deletion strains



Time series after initial perturbation

Modeling



$$\text{Prob}(\text{signal}|\text{point}) = 2\Phi((\text{point} - \text{ref}) / s) - 1$$

Noise models

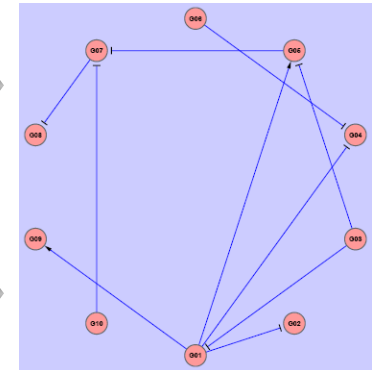
$$\frac{dy_j}{dt} = a_{j0} - a_{jj}y_j + \sum_{k \in S} a_{jk}y_k$$

$$\frac{dy_j}{dt} = \frac{b_{j1}}{1 + \exp\left(a_{j0} + \sum_{k \in S} a_{jk}y_k\right)} - b_{j2}y_j$$

$$\frac{dy_j}{dt} = a_{j0} \prod_{k_1 \in S_1} \left(\frac{b_{jk_1}}{y_{k_1}^{c_{jk_1}} + b_{jk_1}} \right) \prod_{k_2 \in S_2} \left(\frac{y_{k_2}^{c_{jk_2}}}{y_{k_2}^{c_{jk_2}} + b_{jk_2}} \right) - a_{j1}y_j$$

Expression rate models

Predictions



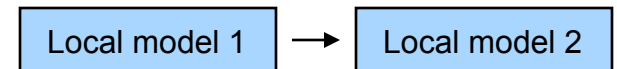
[Yip et al., DREAM3]

Accuracy (AUC)	E. Coli 1	E. Coli 2	Yeast 1	Yeast 2	Yeast 3
Size-10	0.928	0.912	0.949	0.747	0.714
Size-50	0.930	0.924	0.917	0.792	0.805
Size-100	0.948	0.960	0.915	0.856	0.783

Our work: efficiently propagating known information

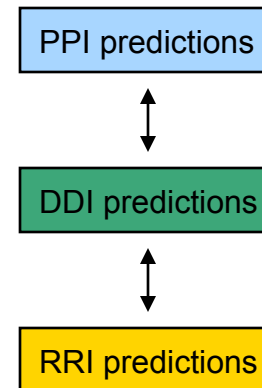
Training set expansion

- Motivation: lack of training examples
- Expand training sets horizontally



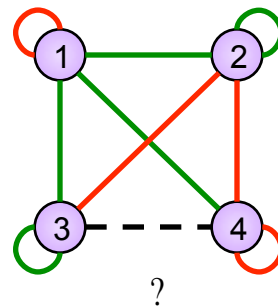
Multi-level learning

- Motivation: hierarchical nature of interaction
- Expand training sets vertically



DREAM3 *in silico* regulatory network reconstruction challenge

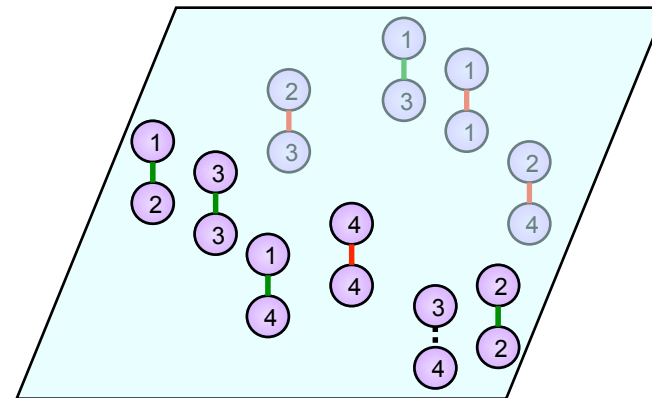
Global vs. local modeling



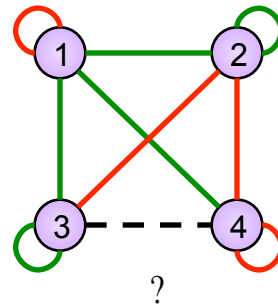
Global modeling: build one model for the whole network

Example - Pairwise kernel:
consider object pairs instead
of individual objects

Problem: $O(n^2)$ instances,
 $O(n^4)$ kernel elements

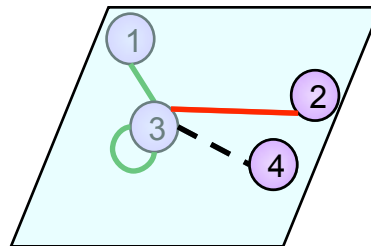


Global vs. local modeling



Local modeling: build one model for each node

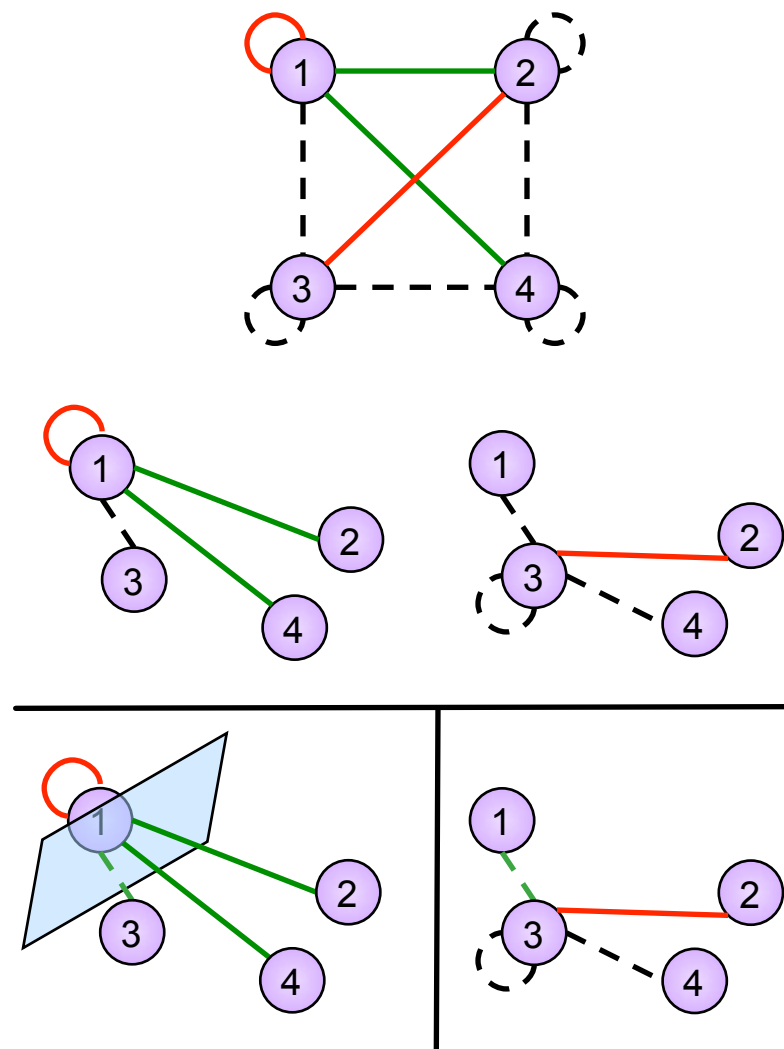
Model for node 3:



Problem: insufficient and unevenly distributed training data (what if node 3 has no known interactions at all?)

Prediction propagation

- Goal: keep the flexibility of local modeling, but tackle the data sparsity problem
- Motivation: some objects have more examples than others
- Our approach:
 - ◊ Learn models for objects with more examples first
 - ◊ Propagate the most confident predictions as auxiliary examples of other objects



Prediction accuracy (AUC)

	phy	loc	exp-gasch	exp-spellman	y2h-ito	y2h-uetz	tap-gavin	tap-krogan	int
Mode 1									
direct	58.04	66.55	64.61	57.41	51.52	52.13	59.37	61.62	70.91
kCCA	65.80	63.86	68.98	65.10	50.89	50.48	57.56	51.85	80.98
kML	63.87	68.10	69.67	68.99	52.76	53.85	60.86	57.69	73.47
em	71.22	75.14	67.53	64.96	55.90	53.13	63.74	68.20	81.65
local	71.67	71.41	72.66	70.63	67.27	67.27	64.60	67.48	75.65
local+pp	73.89	75.25	77.43	75.35	71.60	71.51	74.62	71.39	83.63
local+ki	71.68	71.42	75.89	70.96	69.40	69.05	70.53	72.03	81.74
local+pp+ki	72.40	75.19	77.41	73.81	70.44	70.57	73.59	72.64	83.59

Observations:

- Highest accuracy by training set expansion
- Over fitting of local modeling without training set expansion
- Prediction propagation theoretically related to co-training (Blum and Mitchell, 1998)
 - ◇ Semi-supervised (Similarity with PSI-BLAST)

From horizontal to vertical

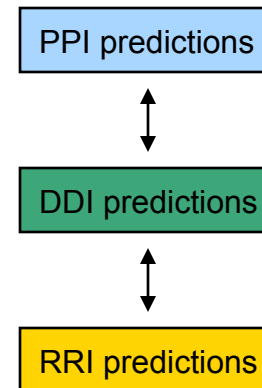
Training set expansion

- Motivation: lack of training examples
- Expand training sets horizontally

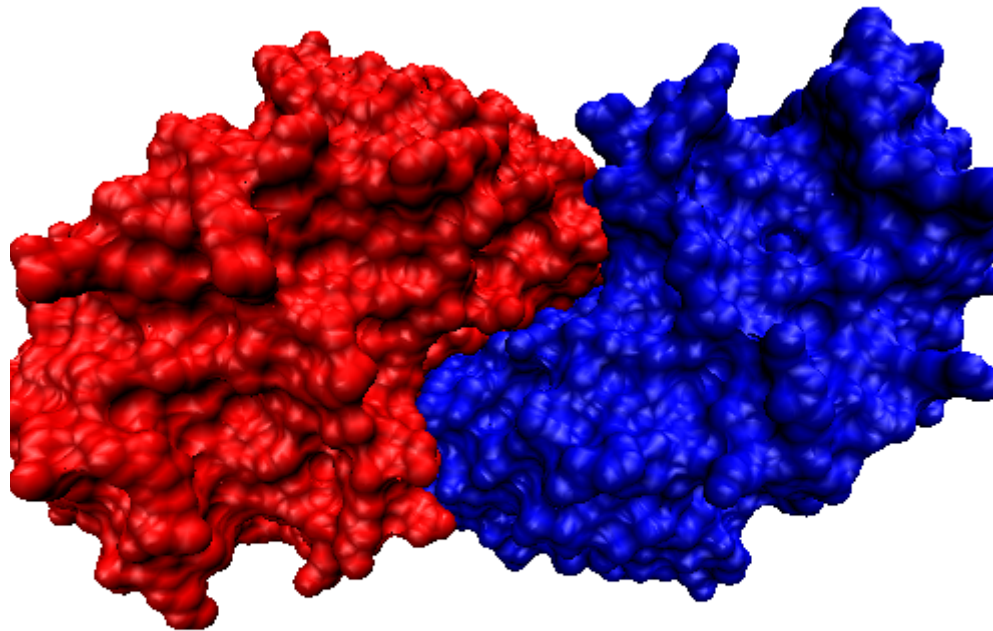


Multi-level learning

- Motivation: hierarchical nature of interaction
- Expand training sets vertically



Protein interaction

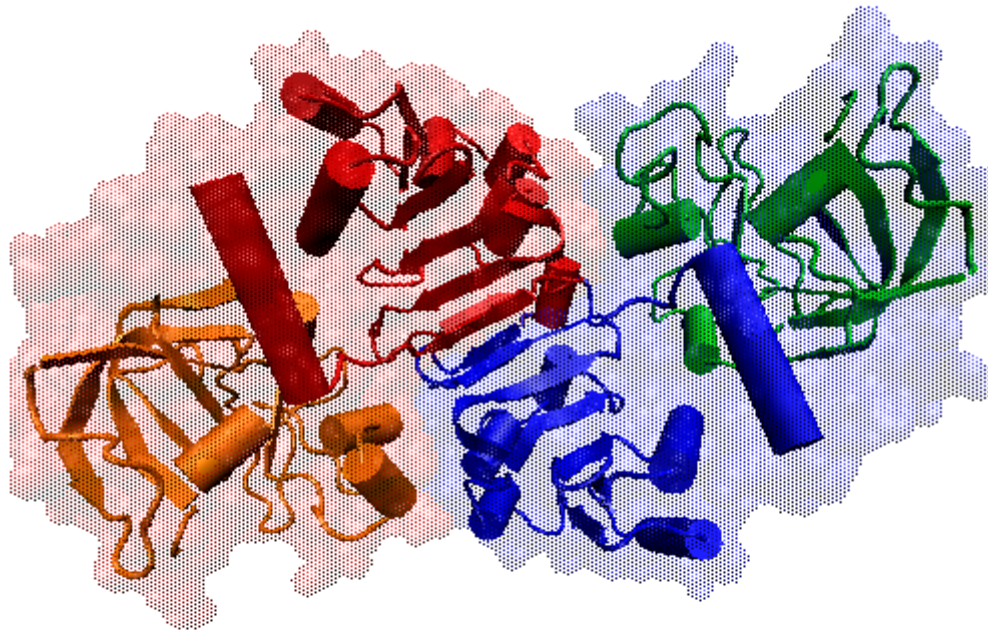


Yeast NADP-dependent alcohol dehydrogenase 6 (PDB: 1piw)

Protein-level features for interaction prediction: functional genomic information

[Yip and Gerstein, in revision]

Domain interaction

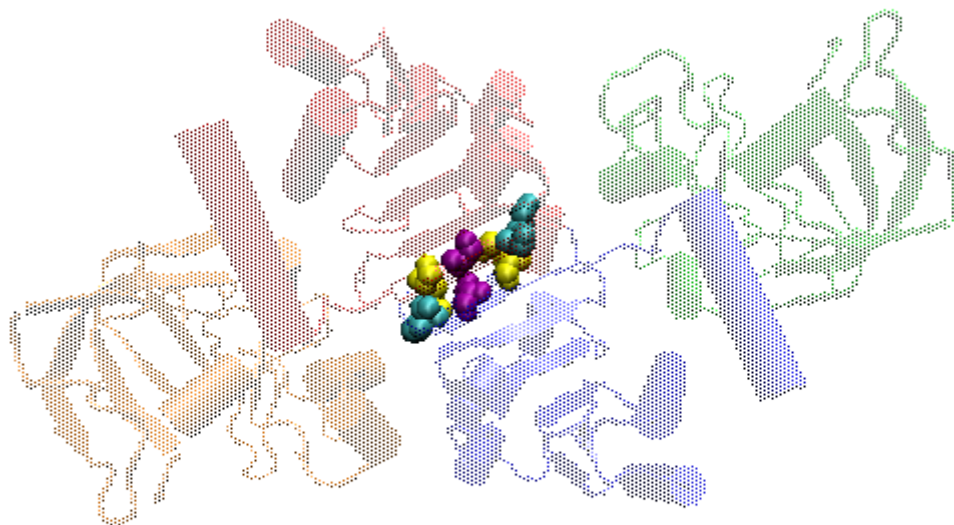


Pfam domains: PF00107 (inner) and PF08240 (outer)

Domain-level features for interaction prediction: evolutionary information

[Yip and Gerstein, in revision]

Residue interaction

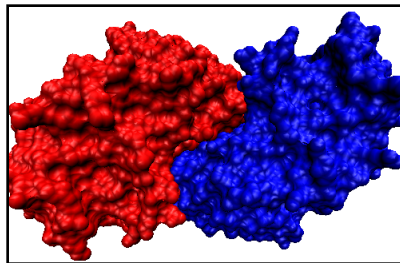


Interacting residues: 283 (yellow) with 287 (cyan), and 285 (purple) with 285

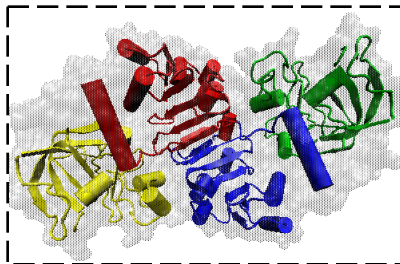
Residue-level features for interaction prediction: physical-chemical information

[Yip and Gerstein, in revision]

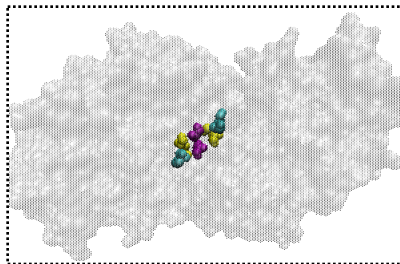
Combining the three problems



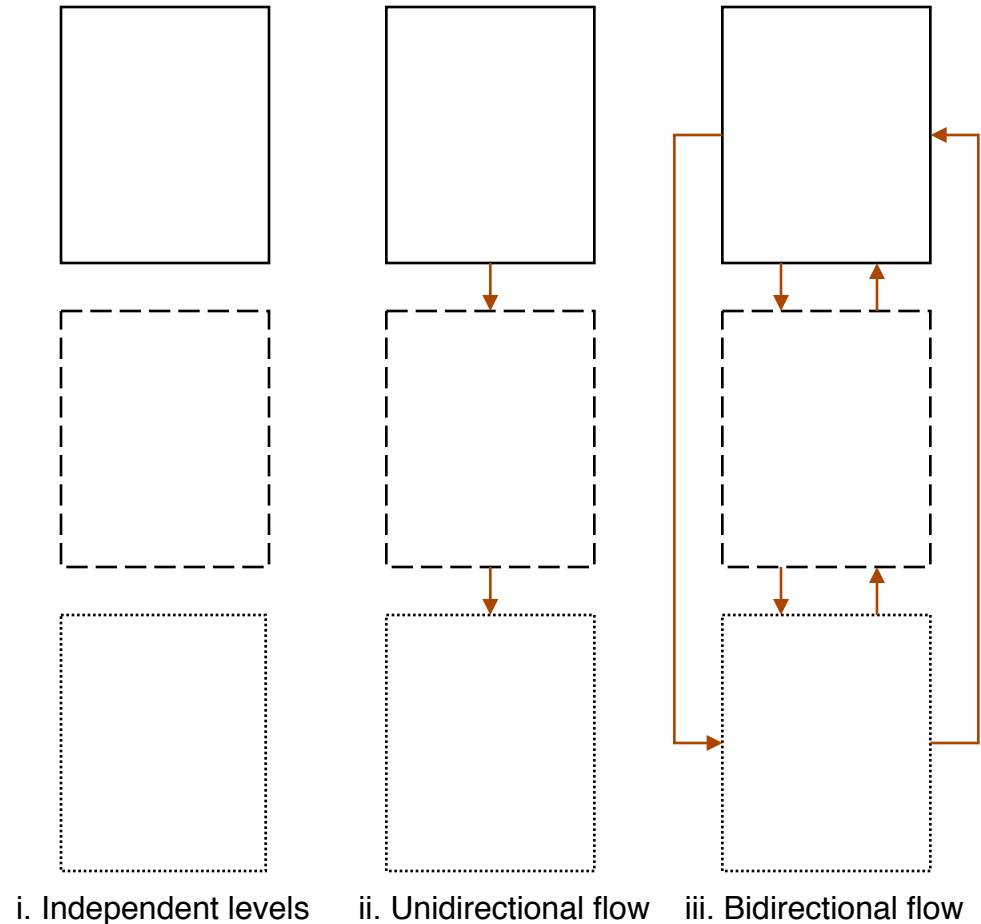
Protein interactions



Domain interactions

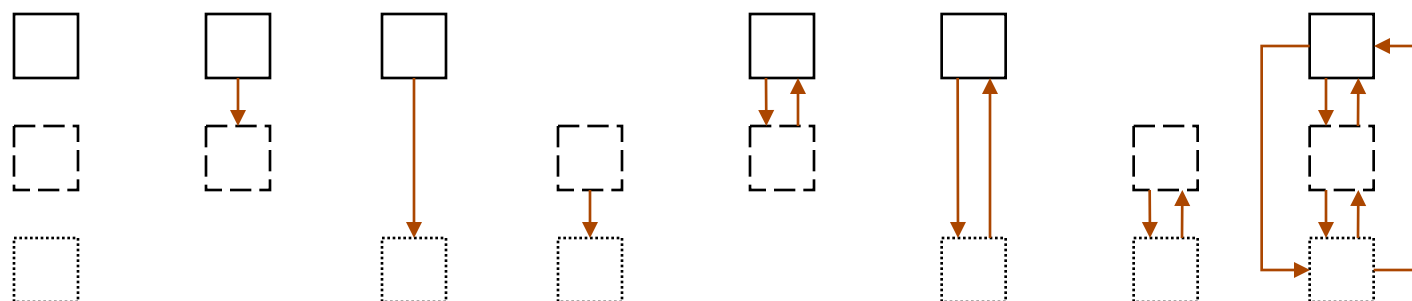


Residue interactions



Empirical results (AUCs)

	Ind. levels	Unidirectional flow			Bidirectional flow			
Level		PD	PR	DR	PD	PR	DR	PDR
Proteins	71.68				72.23	72.50		72.82
Domains	53.18	61.51			71.71		68.94	71.20
Residues	57.36		54.89	53.81		72.26	63.16	77.86



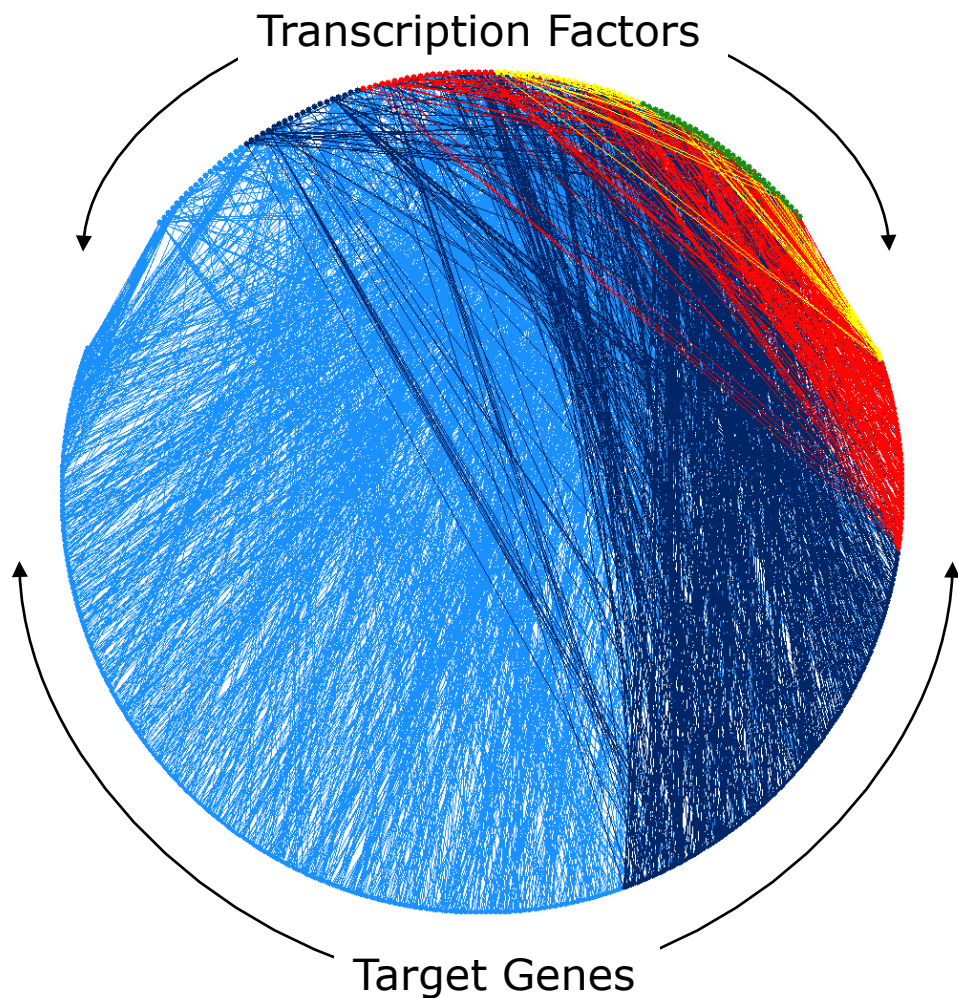
- Highest accuracy by bidirectional flow
- Additive effect: 2 vs. 3 levels

Network Dynamics #1: Cellular States

How do networks change across different cellular states?
How can this be used to assign function to a protein?



Dynamic Yeast TF network



- Analyzed network as a static entity
- But network is *dynamic*
 - ◇ Different sections of the network are active under different cellular conditions
- Integrate gene expression data

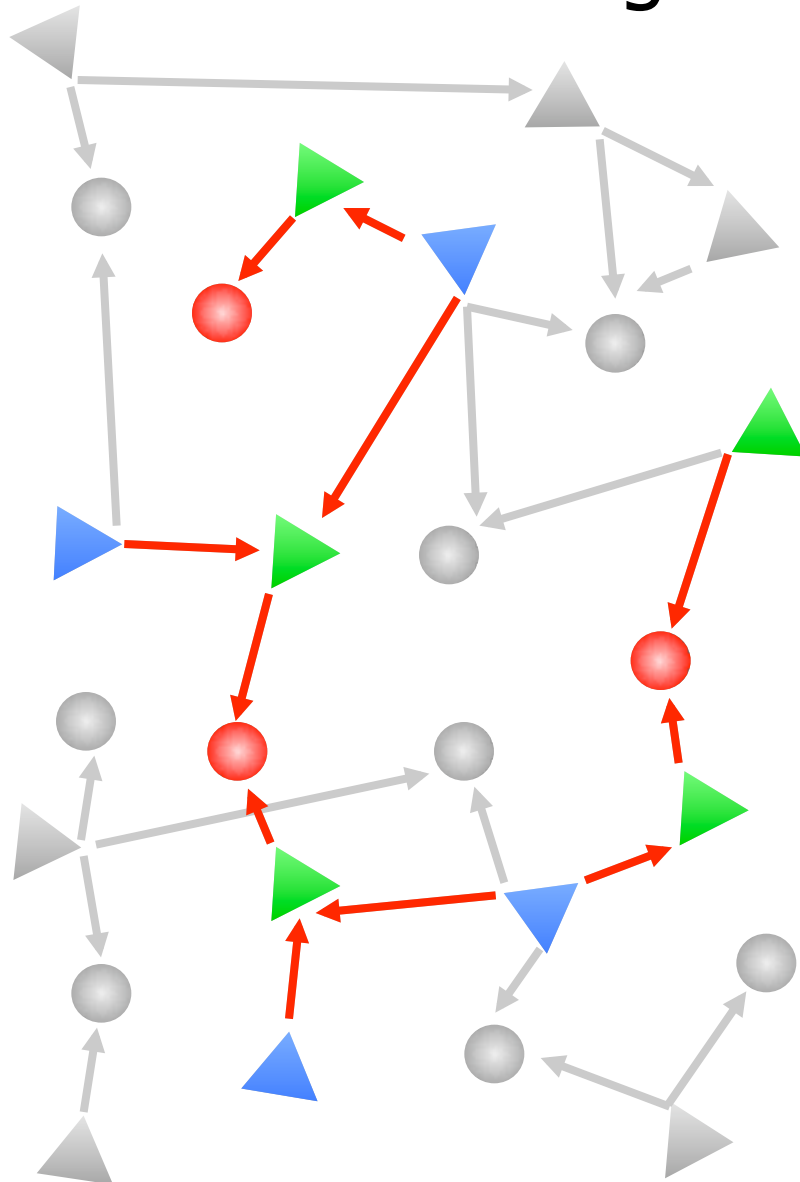
Luscombe et al. Nature 431: 308

Gene expression data for five cellular conditions in yeast

Cellular condition	
Multi-stage	Cell cycle
	Sporulation
Binary	Diauxic shift
	DNA damage
	Stress response

[Brown, Botstein, Davis....]

Backtracking to find active sub-network

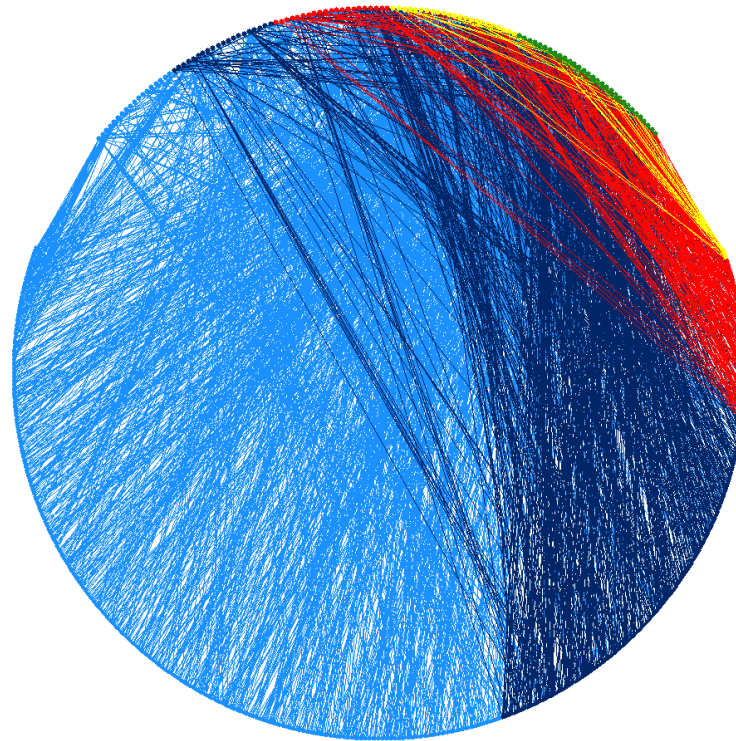


- Define differentially expressed genes
- Identify TFs that regulate these genes
- Identify further TFs that regulate these TFs

Active regulatory sub-network

Network usage under different conditions

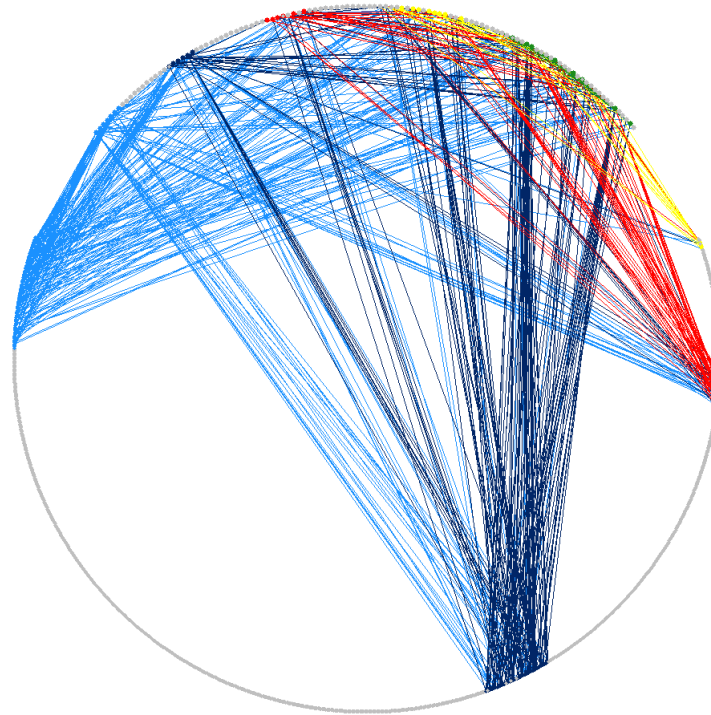
static



Luscombe et al. Nature 431: 308

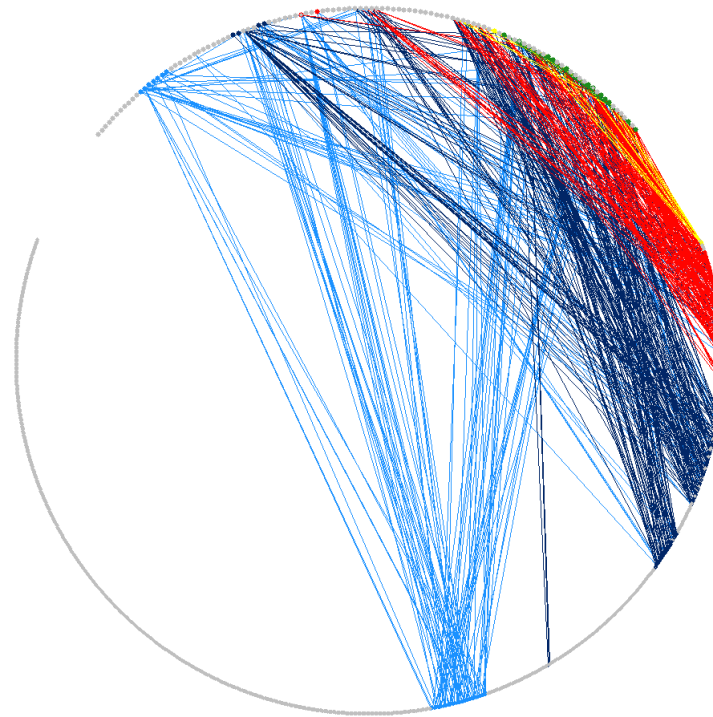
Network usage under different conditions

cell cycle



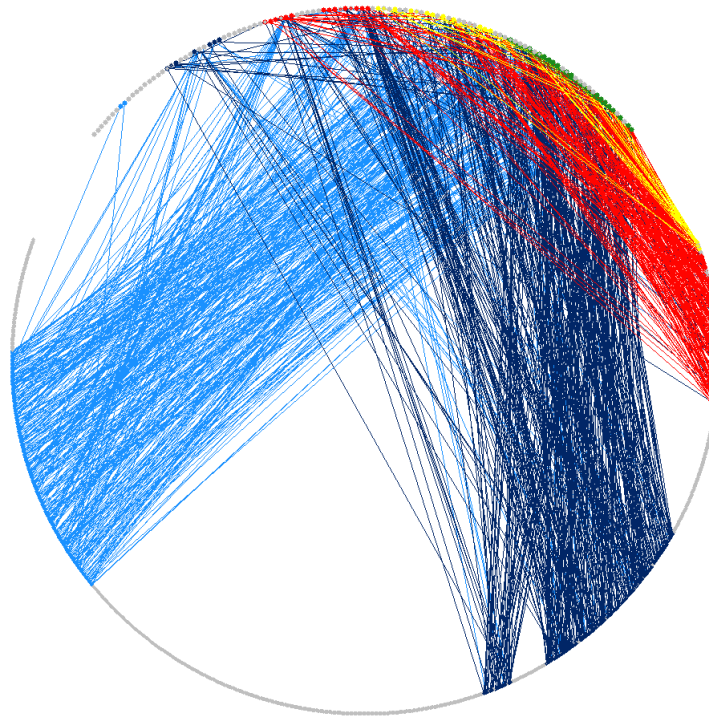
Network usage under different conditions

sporulation



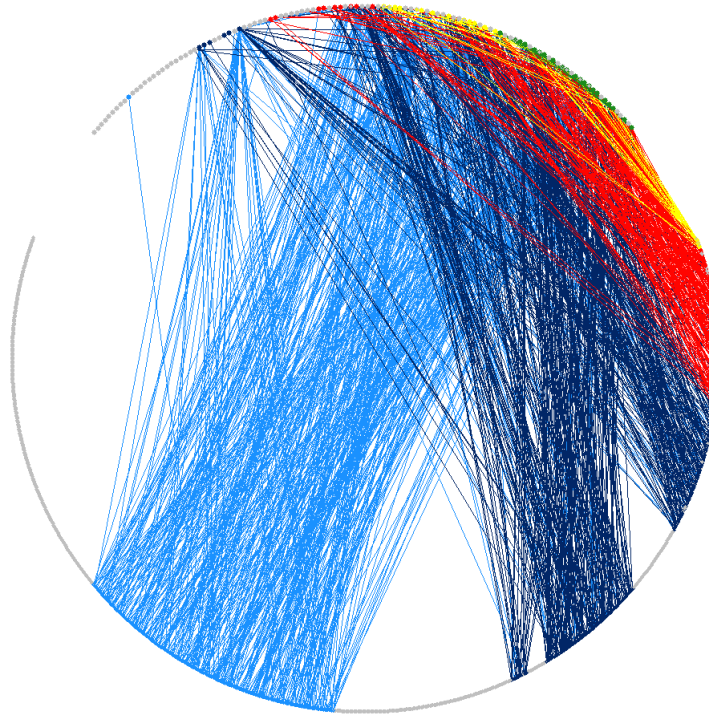
Network usage under different conditions

diauxic shift



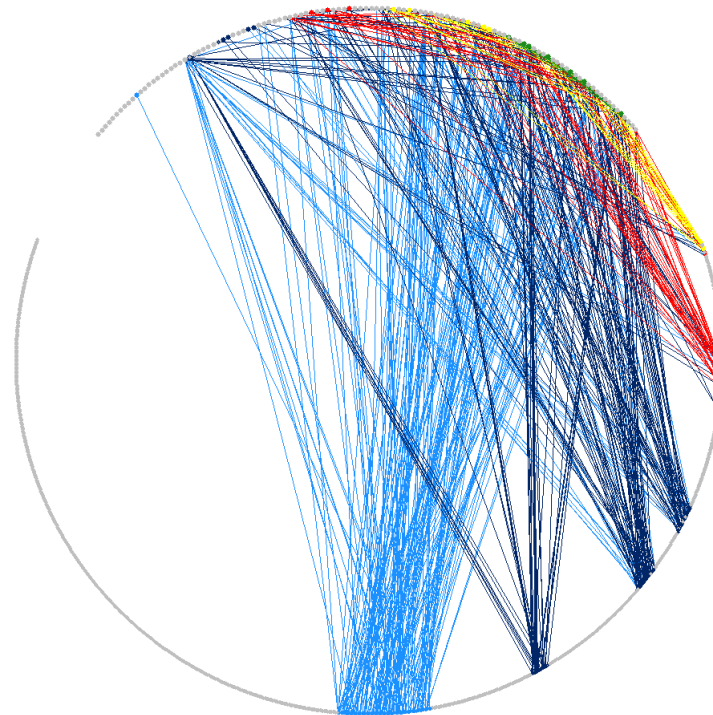
Network usage under different conditions

DNA damage



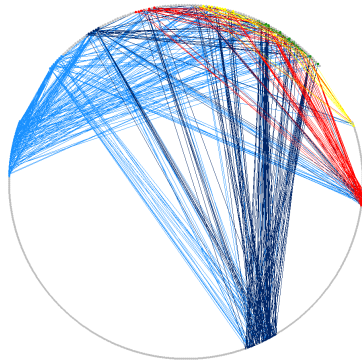
Network usage under different conditions

stress response

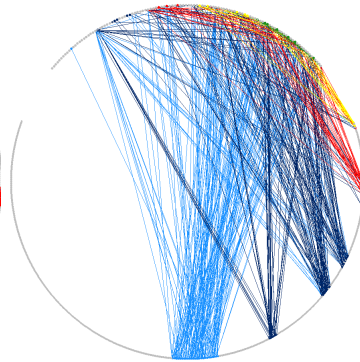


Network usage under different conditions

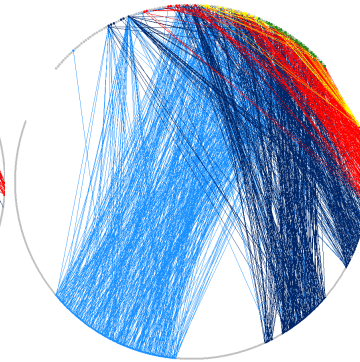
Cell cycle



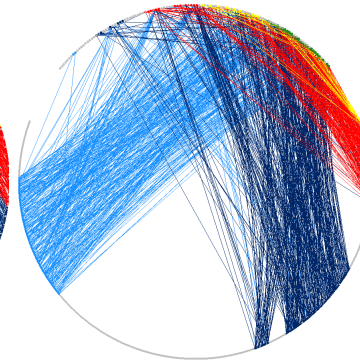
Sporulation



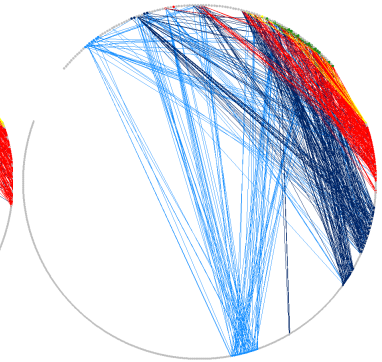
Diauxic shift



DNA damage



Stress



SANDY:

1. Standard graph-theoretic statistics:

- Global topological measures
- Local network motifs

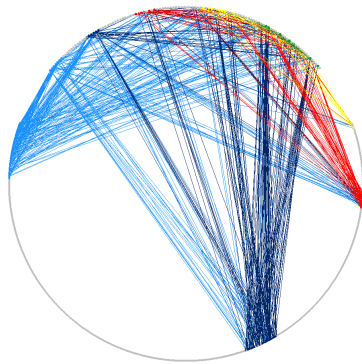
2. Newly derived follow-on statistics:

- Hub usage
- Interaction rewiring

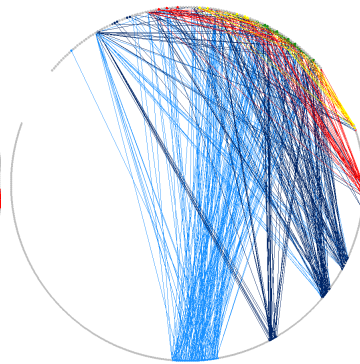
3. Statistical validation of results

Network usage under different conditions

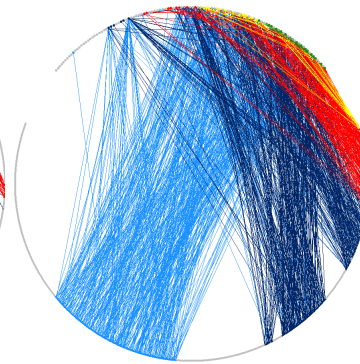
Cell cycle



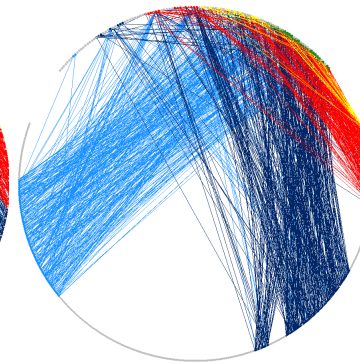
Sporulation



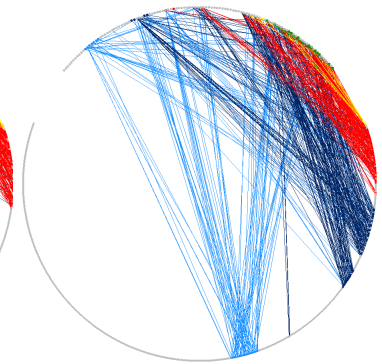
Diauxic shift



DNA damage



Stress



SANDY:

1. Standard graph-theoretic statistics:

- Global topological measures
- Local network motifs

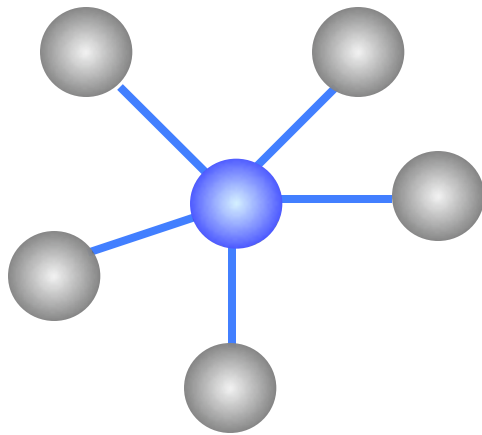
2. Newly derived follow-on statistics:

- Hub usage
- Interaction rewiring

3. Statistical validation of results

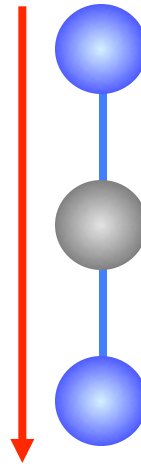
Global topological measures

Indicate the gross topological structure of the network



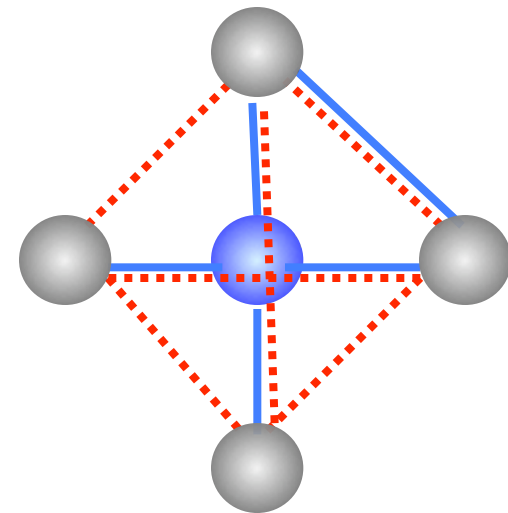
Degree (K)

5



Path length (L)

2



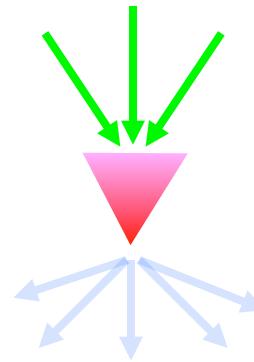
Clustering coefficient (C)

1/6

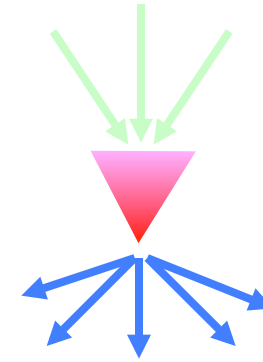
Interaction and expression networks are ***undirected***

[Barabasi]

Global topological measures for directed networks



In-degree
3

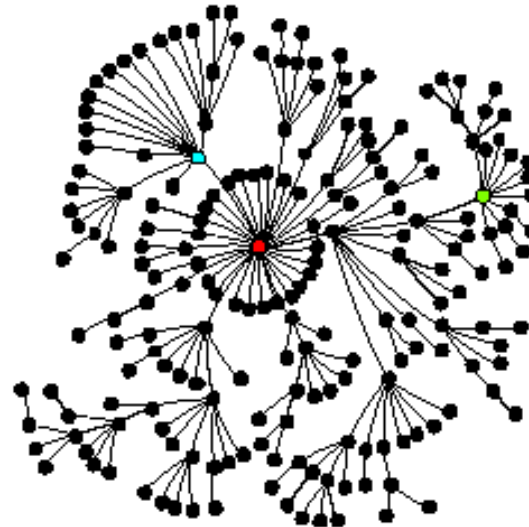
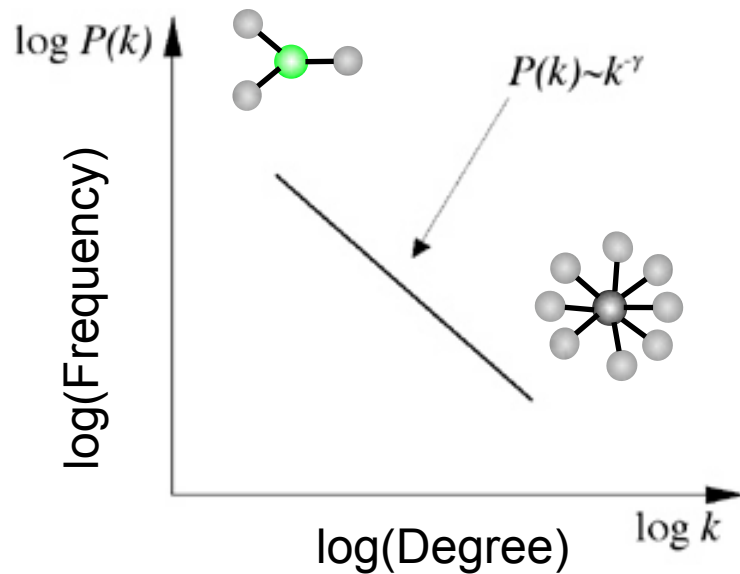


Out-degree
5

Regulatory and metabolic networks are ***directed***

Scale-free networks

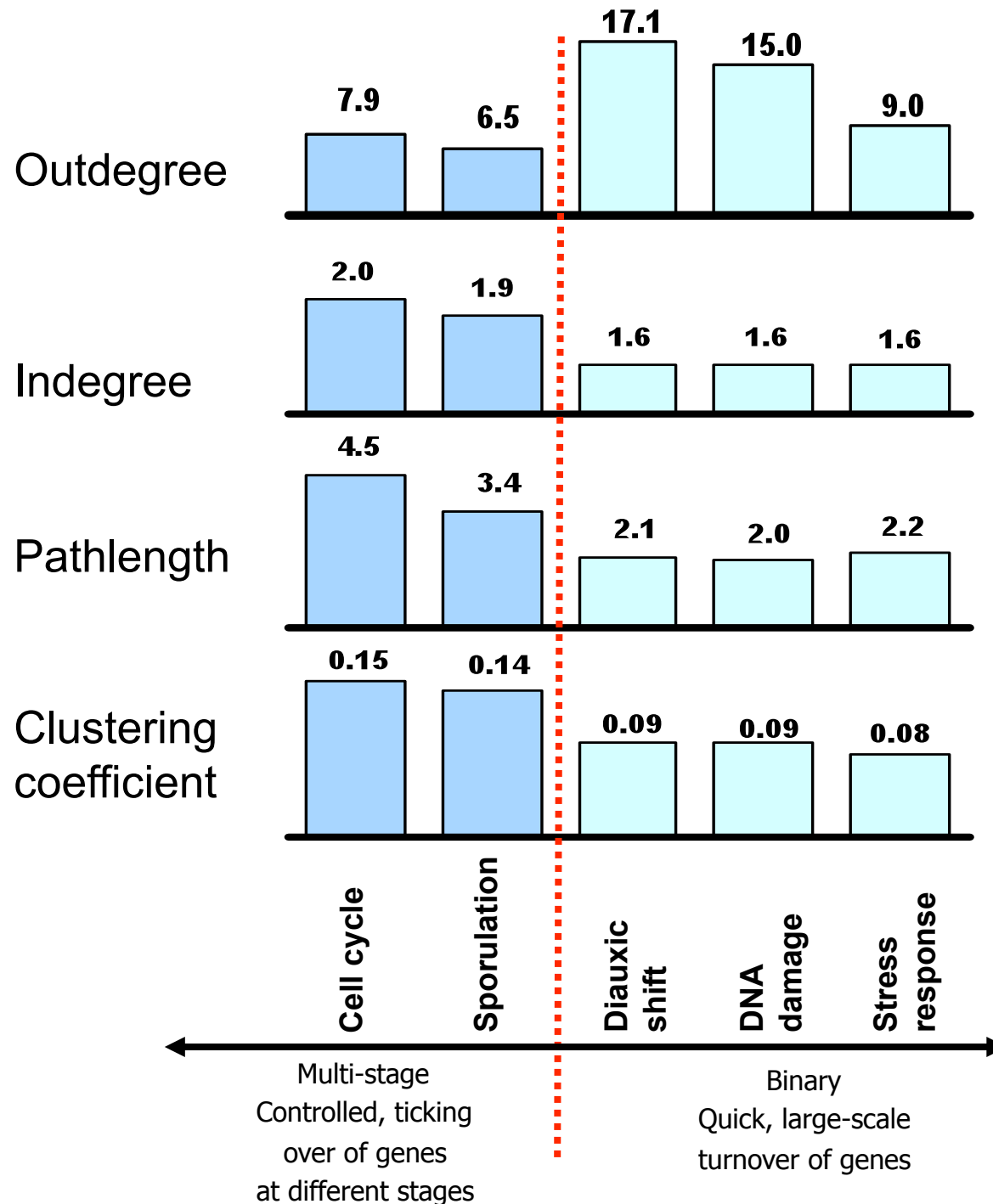
Power-law distribution



Hubs dictate the structure of the network

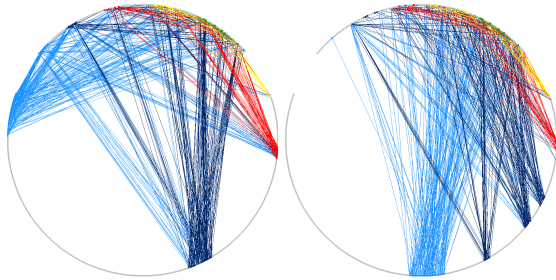
[Barabasi]

Analysis of condition-specific subnetworks in terms of global topological statistics

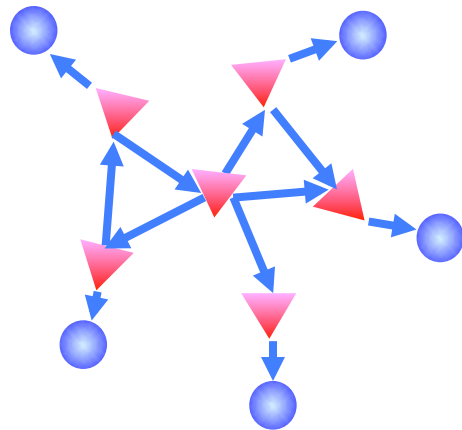


Luscombe et al. Nature 431: 308

Cell cycle **Sporulation**

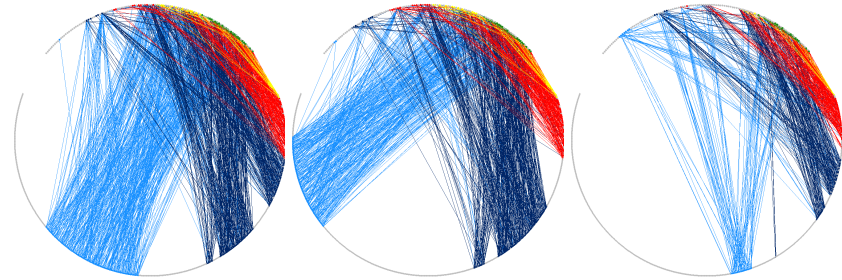


multi-stage conditions

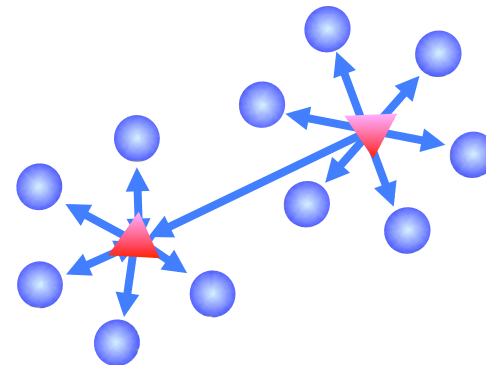


less pronounced
longer
more
complex loops (FFLs)

Diauxic shift **DNA damage** **Stress**



binary conditions

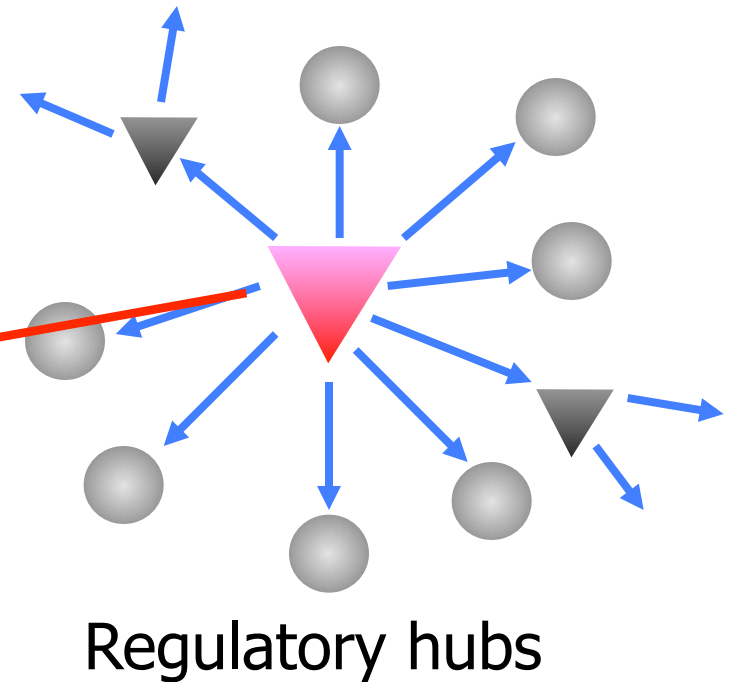
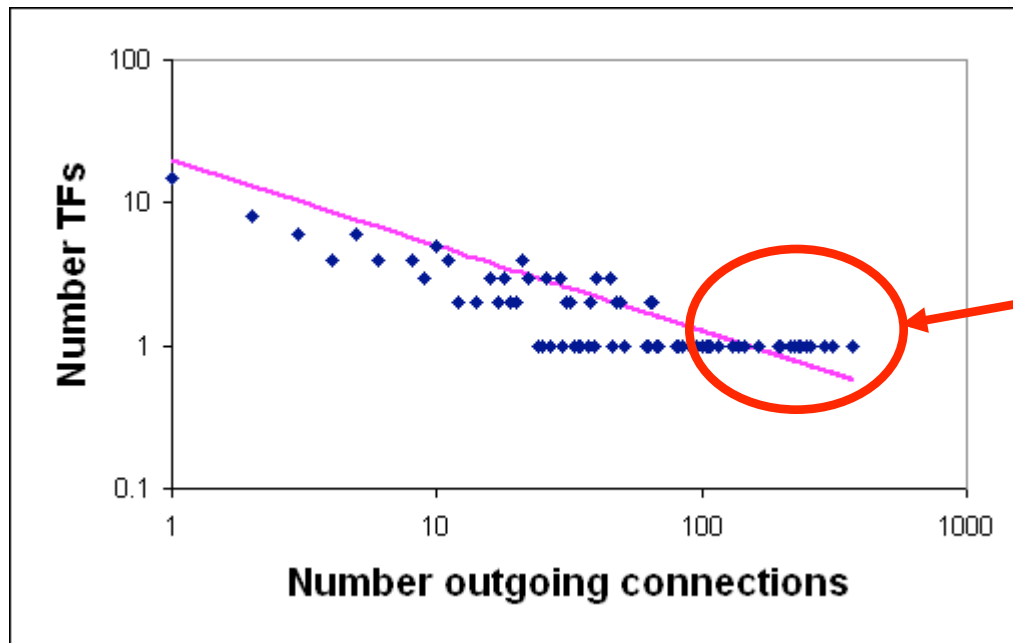


more pronounced
shorter
less
simpler (SIMs)

Summary

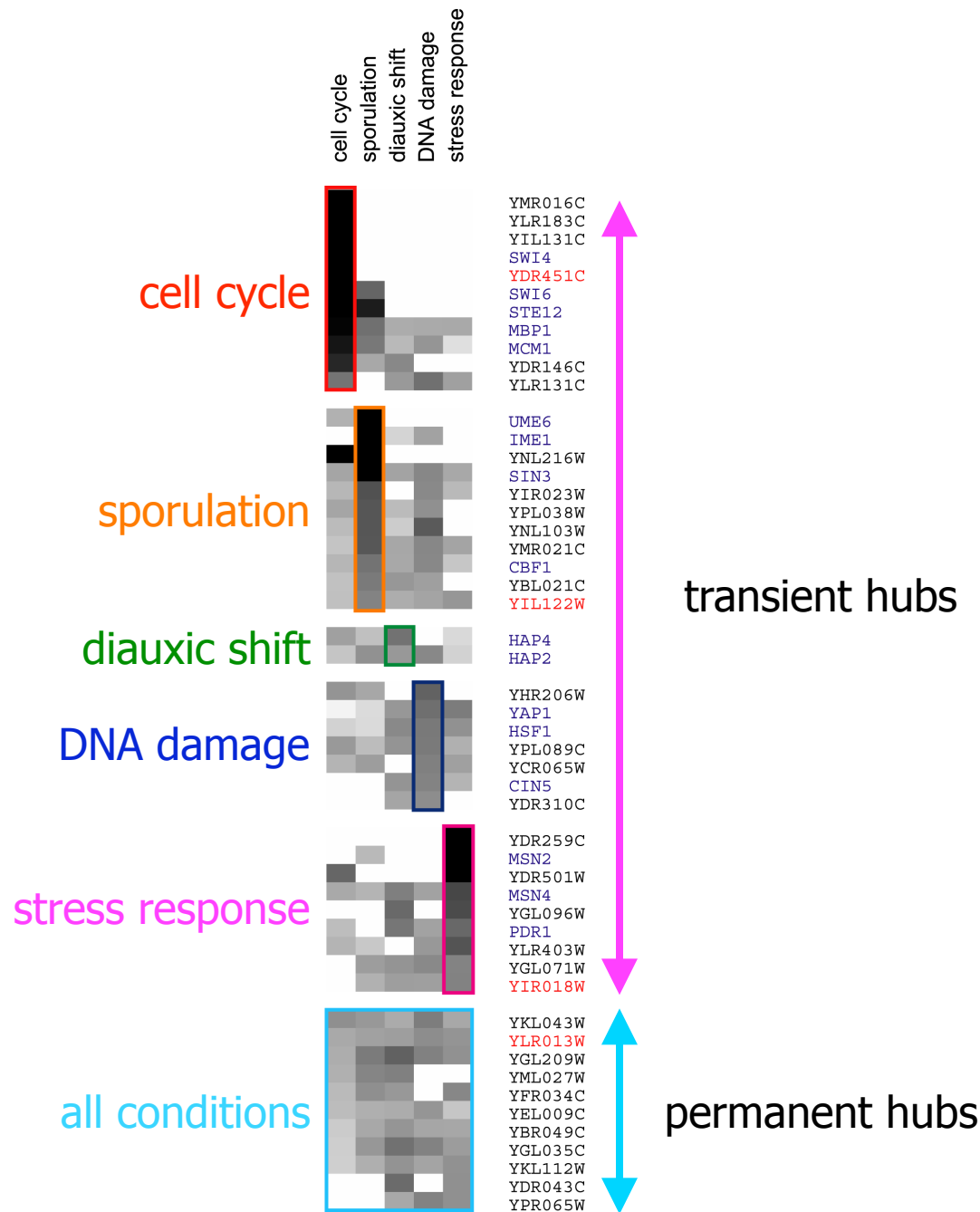
Hubs
Path Lengths
TF inter-regulation
Motifs

Transient Hubs



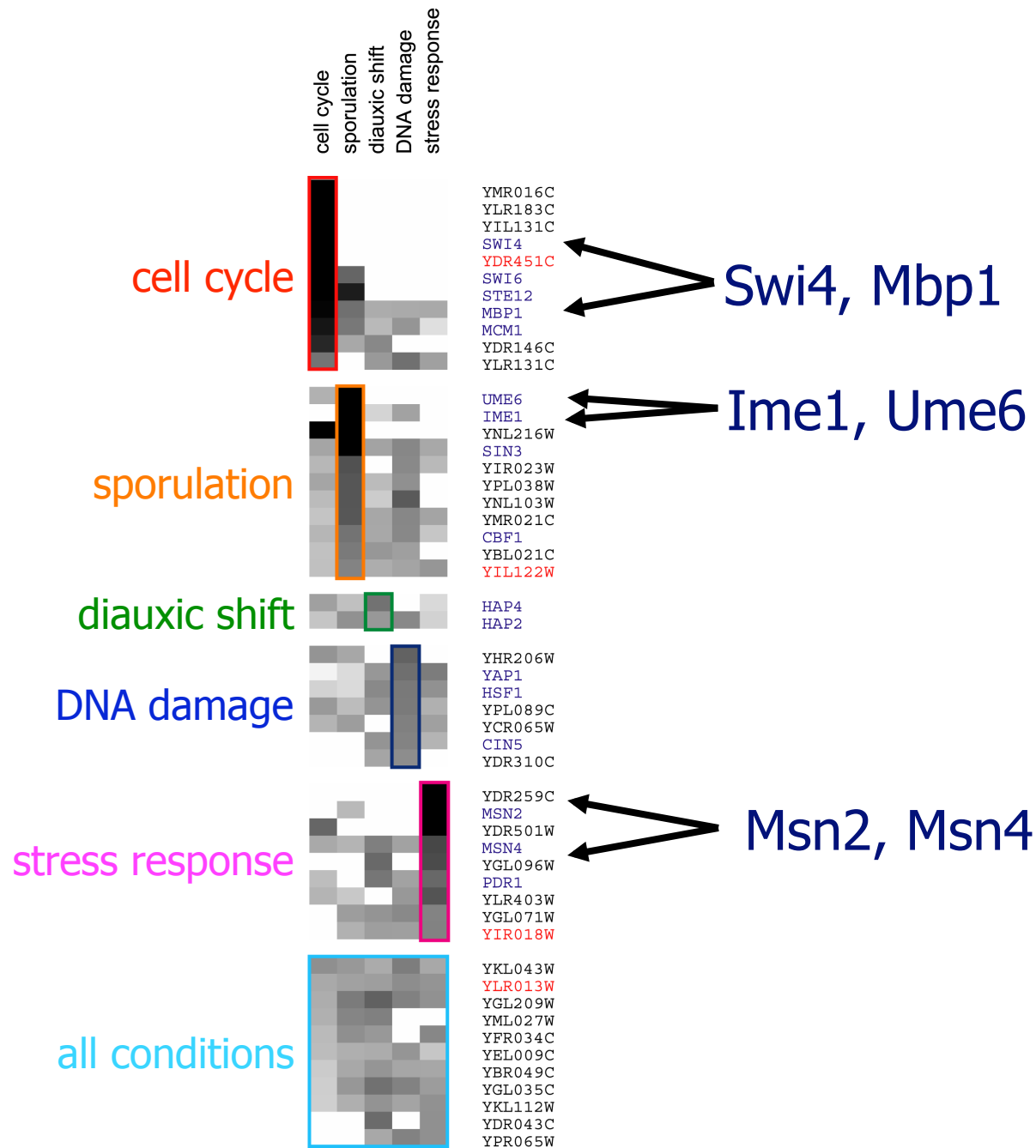
- Questions:
 - ◇ Do hubs stay the same or do they change over between conditions?
 - ◇ Do different TFs become important?
- Our Expectations
 - ◇ Literature:
 - Hubs are permanent features of the network regardless of condition
 - ◇ Random networks (sampled from complete regulatory network)
 - Random networks converge on same TFs
 - 76-97% overlap in TFs classified as hubs (*ie* hubs are permanent)

Luscombe et al. Nature 431: 308

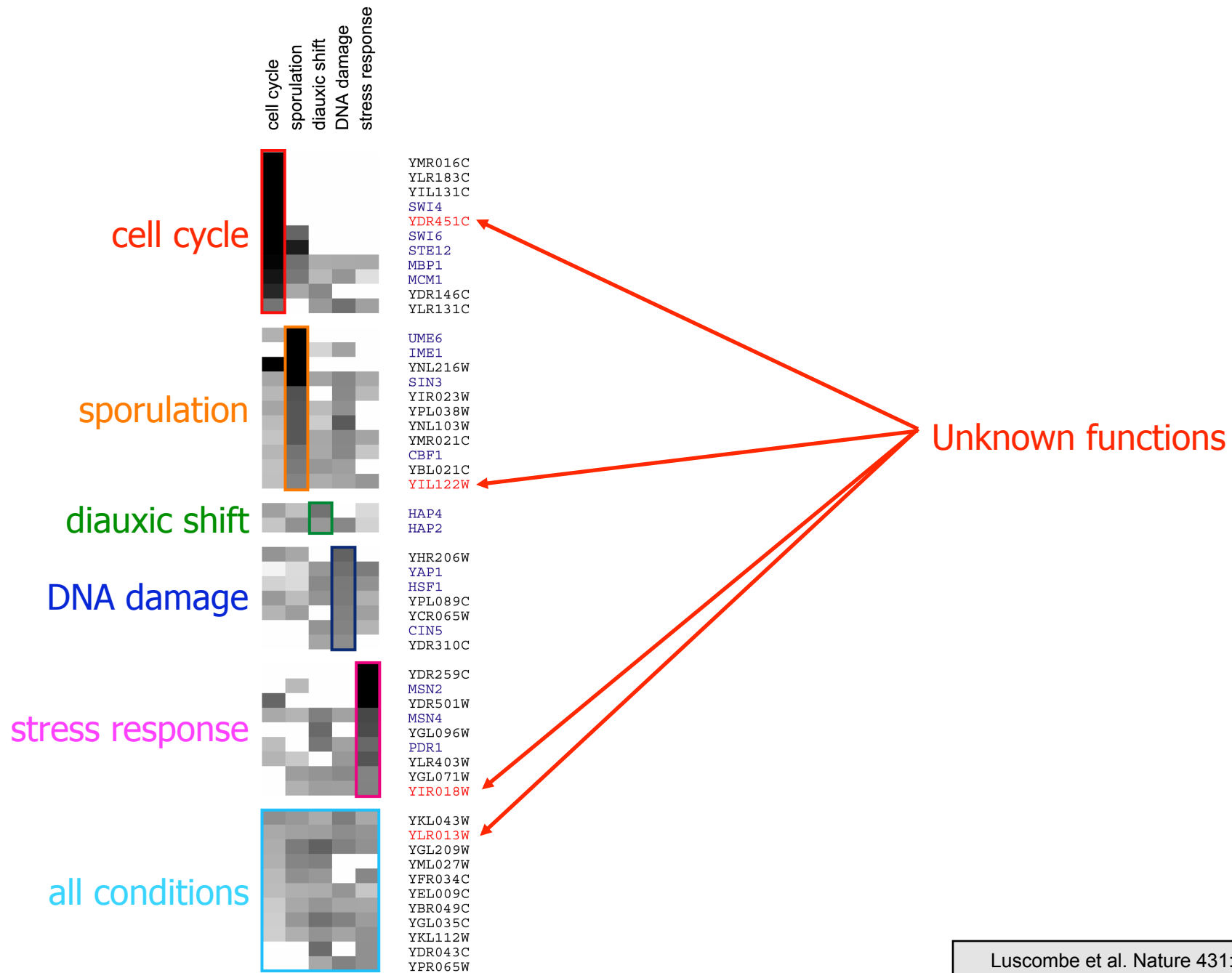


- Some permanent hubs
 - ◊ house-keeping functions
- Most are transient hubs
 - ◊ Different TFs become key regulators in the network
- Implications for condition-dependent vulnerability of network

Luscombe et al. Nature 431: 308



Luscombe et al. Nature 431: 308



Luscombe et al. Nature 431: 308

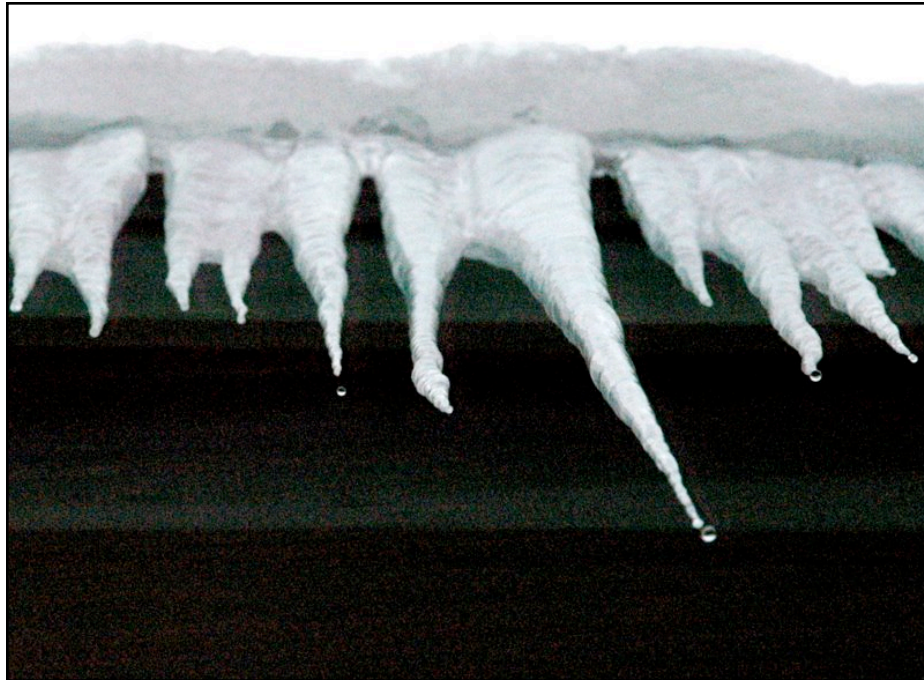
Network Dynamics #2:

Environments

How do molecular networks change across environments?

What pathways are used more ?

Used as a biosensor ?



What is metagenomics?

Genomics Approach

Culture Microbes



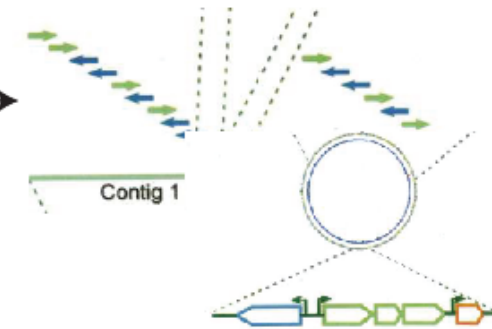
Extract DNA



Sequence

```
ATCGTATA
CGCGAAG
ACGTCTGA
AGTGCTGCT
```

Assemble and Annotate



PROBLEM: Estimated that less than 1% can be cultured in the lab

Metagenomics Approach

Collect Sample



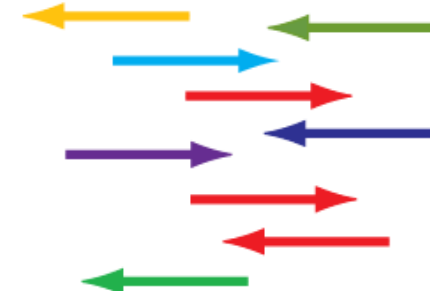
Extract DNA



Sequence

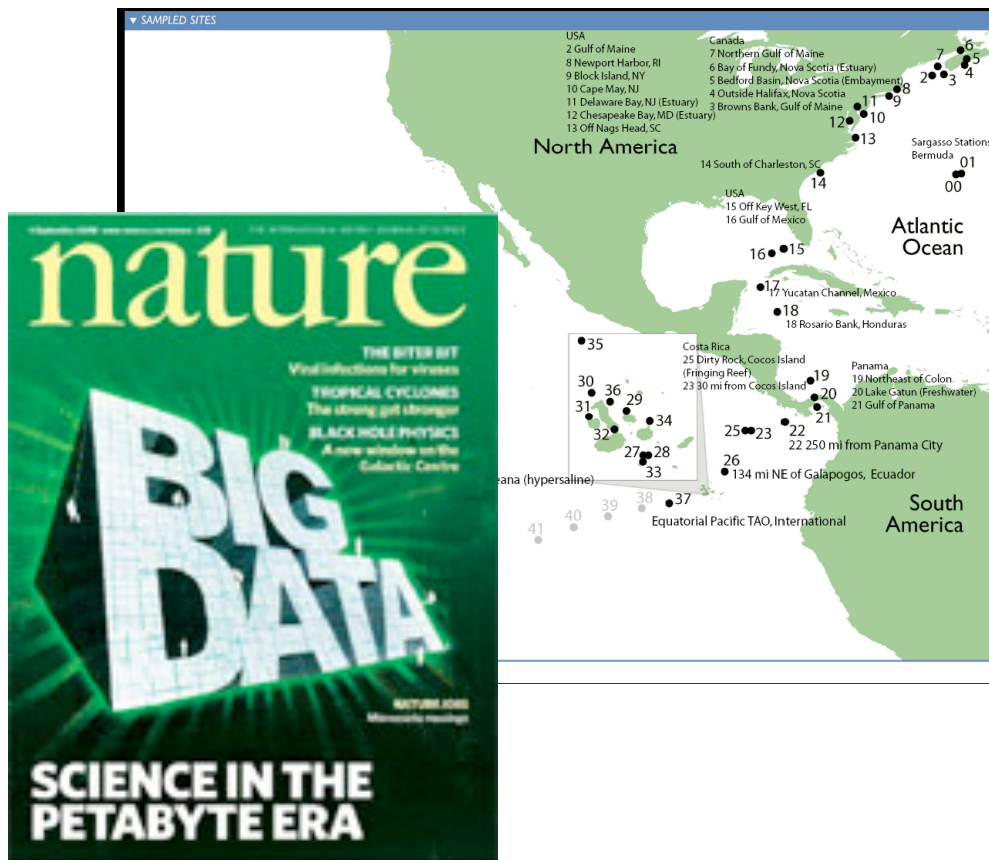
```
ATCGTGATAGATGATAGTAGA
ATGCTGCATGCATCTAGCACT
ACAGTAGCTAGCTACGTAATA
CAGCTGACTAGCTAGCTAGCT
ACGTAGCATGCTAGCTAGCAG
ACGTACGTAGCTAGCTAGTAG
ACGTACGTACGTAGCTAGCATC
AGTCGACTGAGCCAGTGATGAT
ACGATGCATGAGCAGATGCTAC
AGATCGTAGCATGCTAGCATGCT
ACGTACGTAGCTAGCTAGCTAAG
AGCTAGCATGCTAGTAGCATGAG
ACGATGCTAGCTAGCTAGCTGATA
TCGATCAGCATGCTACGATGCAAG
ACGATCGATGCTAGCTAGCAT
AGCTAGCTAGTCAGCTAGCTAGTG
```

Partially Assemble and Annotate



PROBLEM: Lose information about which gene belongs to which microbe.

Global Ocean Survey Statistics (GOS)



6.25 GB of data
7.7M Reads
1 million CPU hours
to process

Rusch, et al., PLOS Biology 2007

Pathway Sequences (Community Function)

Environmental Features

Metabolic Pathways

	P1	P2	P3		
Sites B1	3800	1400	1000		
B2	2200	100	400		
↓	---	---	---		



Environmental

Metadata	Temp	NaCl	Depth		
Sites B1	15°C	27.2	10 m		
B2	23°C	36.6	5 m		
↓	---	---	---		

READS → PROTEIN FAMILIES → PATHWAYS

CCGTGAGCACGATGCGC-----
 ATGCTCATGCT-----
 ATCGTGACGCGATGC-----
 CCGTGAGCACGATGCGC-----
 ATGCTCATGCT-----
 ATCGTGACGCGATGC-----
 ATGCTCATGCT-----
 GCGATCGATCGATCGTAGC-----
 TGCTGCTAGCATGCT-----
 GCGATCGATCGATCGTAGC-----
 TGCTGCTAGCATGCT-----
 CCGTGAGCACGATGCGC-----
 GTATCGTAGCATGCTT-----
 CCGTGAGCACGATGCGC-----
 GCGATCGATCGATCGTAGC-----



$$P_1 = f_1 + f_2 + f_3$$

$$P_2 = f_4 + f_5 + f_6$$

PATHWAYS

SITES	$P_{1,1} = 2 + 1 + 3$ $P_{1,2} = 5 + 2 + 6$	$P_{2,1} = 2 + 4 + 3$ $P_{2,1} = 5 + 7 + 6$

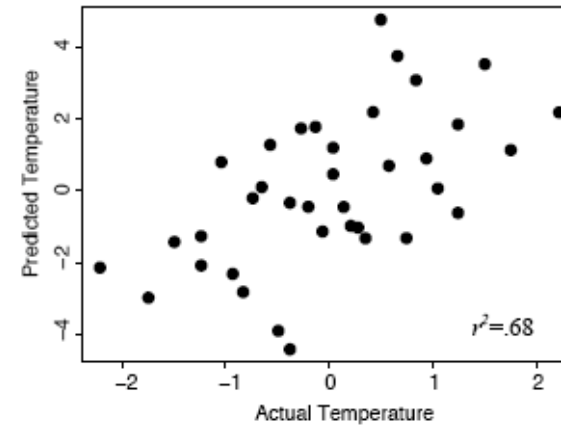
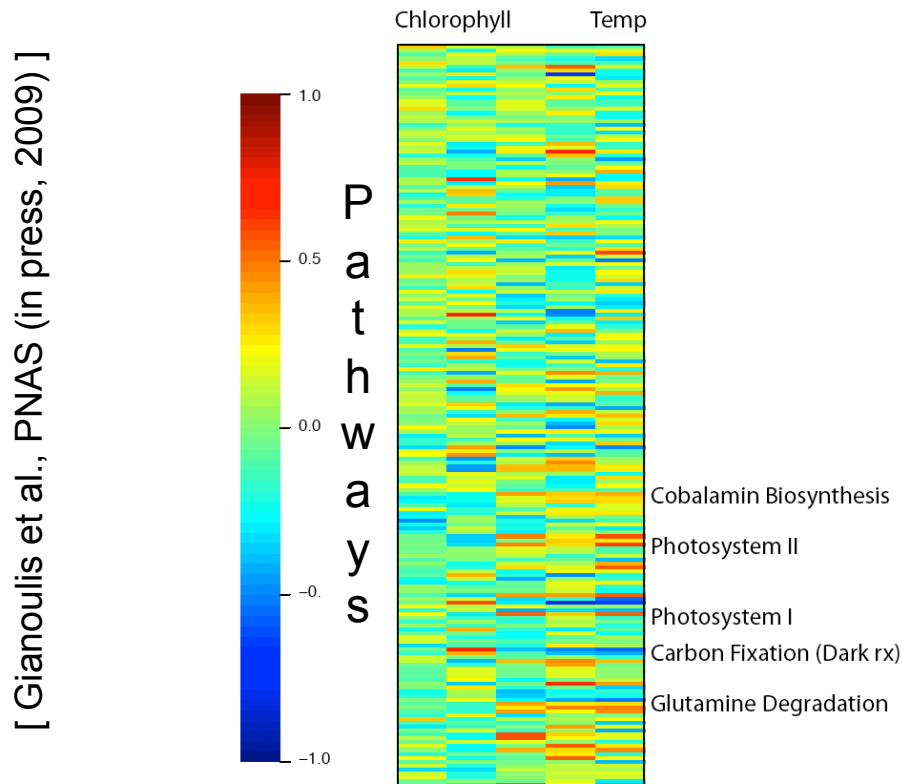
Expressing
data as
matrices
indexed by
site, env. var.,
and pathway
usage

[Rusch et. al., (2007) PLOS Biology;
 Gianoulis et al., PNAS (in press, 2009)]


Simple Relationships: Pairwise Correlations





Environmental Features



Canonical Correlation Analysis: Simultaneous weighting

Score	# of papers published
GRE	

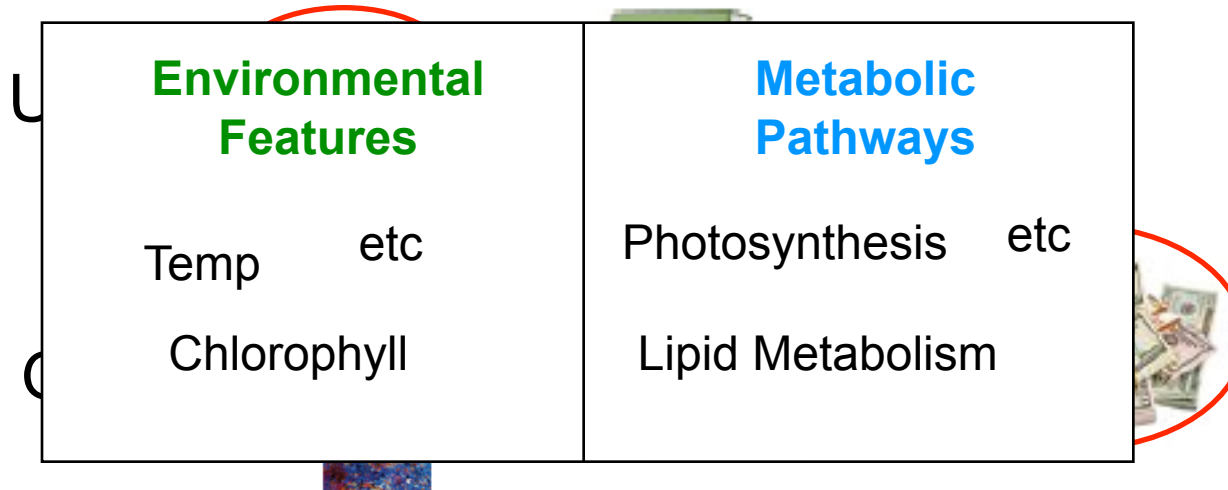
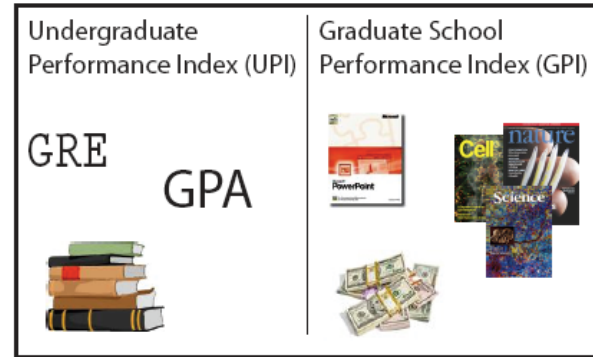
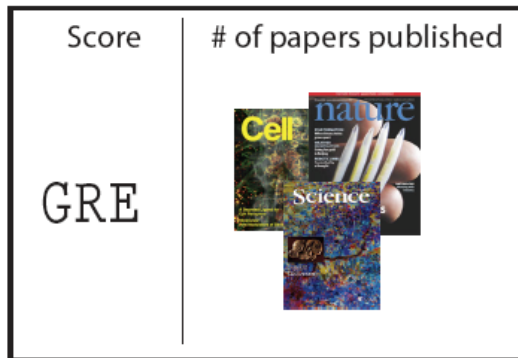
Undergraduate Performance Index (UPI)	Graduate School Performance Index (GPI)
GRE 	

$$\text{UPI} = a \text{ GRE} + b \text{ GPA}$$

$$\text{GPI} = a' \text{ } + b' \text{ } + c' \text{ }$$

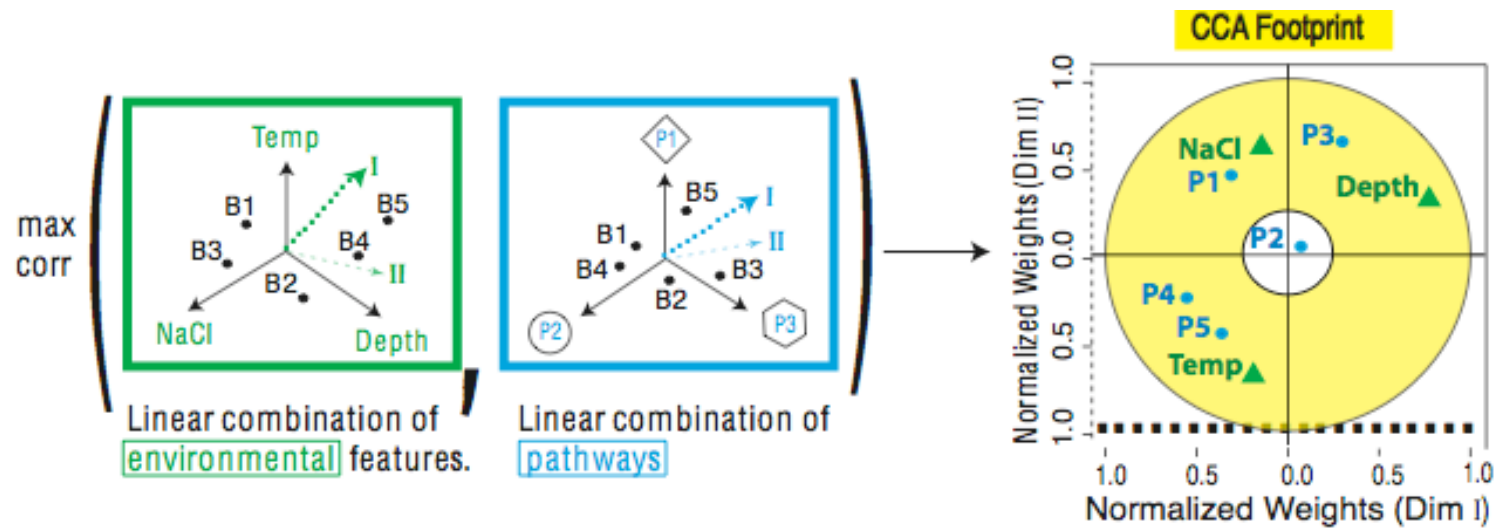
[Gianoulis et al., PNAS (in press, 2009)]

Canonical Correlation Analysis: Simultaneous weighting



[Gianoulis et al., PNAS (in press, 2009)]

Environmental-Metabolic Space



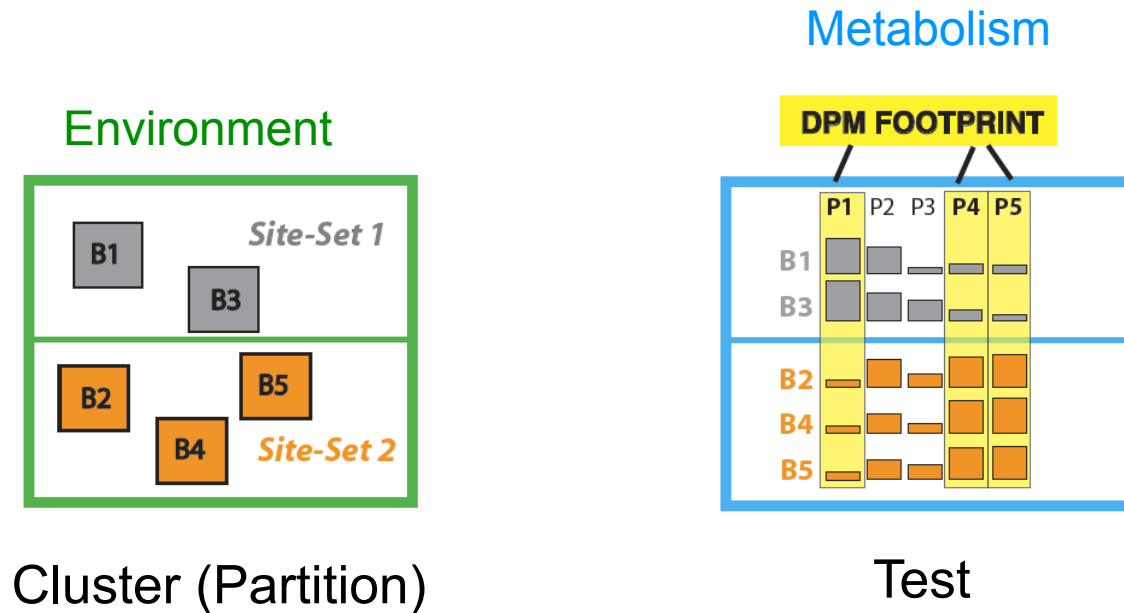
The goal of this technique is to interpret cross-variance matrices
We do this by defining a change of basis.

Given $X = \{x_1, x_2, \dots, x_n\}$ and $Y = \{y_1, y_2, \dots, y_m\}$

$$C = \begin{matrix} \sum_X & \sum_{X,Y} \\ \sum_Y & \sum_{Y,X} \end{matrix} \quad \max_{a,b} \text{Corr}(U,V) = \frac{a' \sum_{12} b}{\sqrt{a' \sum_{11} a} \sqrt{b' \sum_{22} b}}$$

[Gianoulis et al., PNAS (in press, 2009)]

DPM: Discriminative Partition Matching

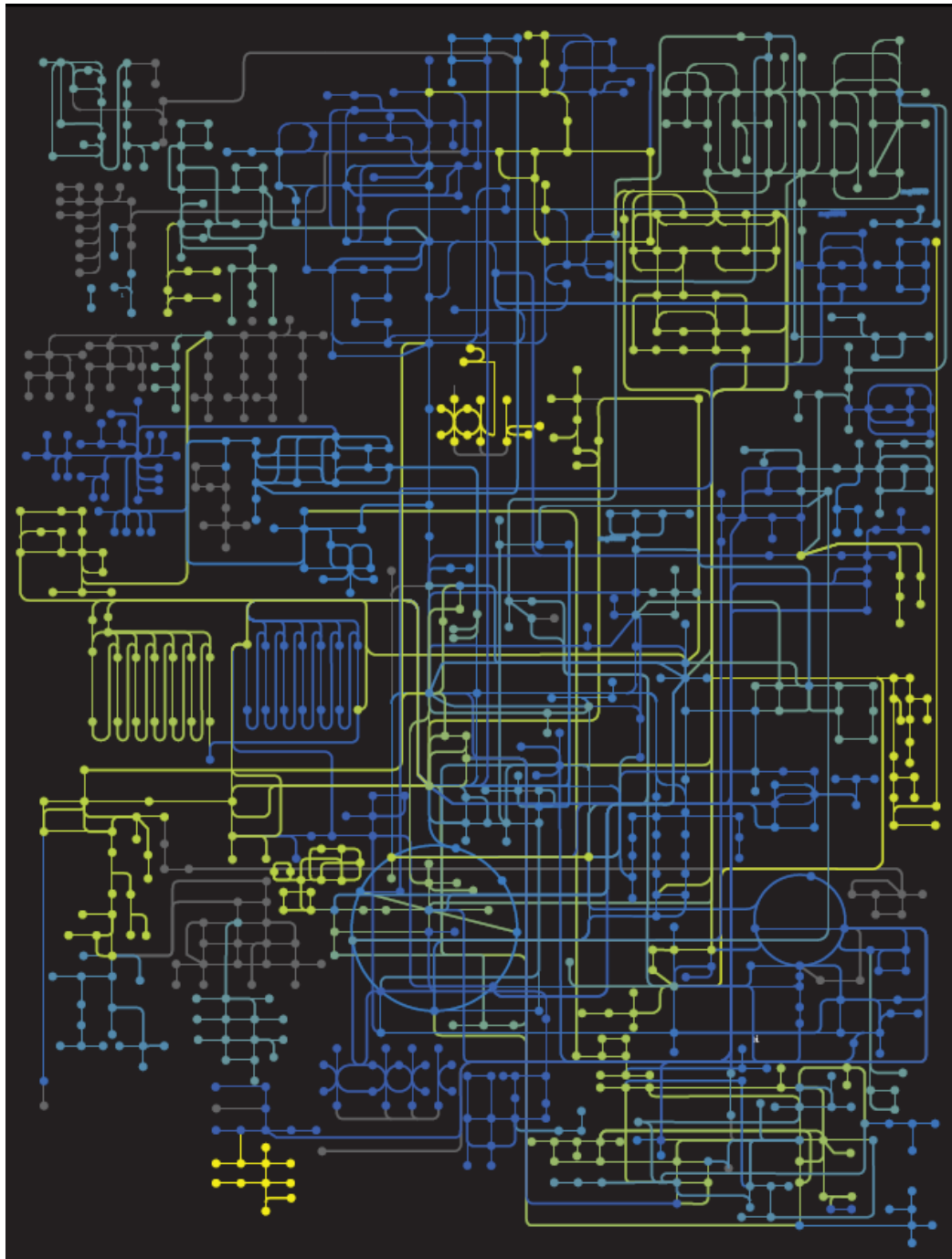


Taurine biosynthesis
 Heme biosynthesis
 Asparagine degradation
 Nitrogen fixation
 Acylglycerol degradation
 Asparagine biosynthesis
 Cysteine Metabolism

Functional class **pval**

InfoStorage & Processing	.07
Cellular Process	.08
Metabolism	4x10 ⁻¹⁴

[Gianoulis et al., PNAS (in press, 2009)]

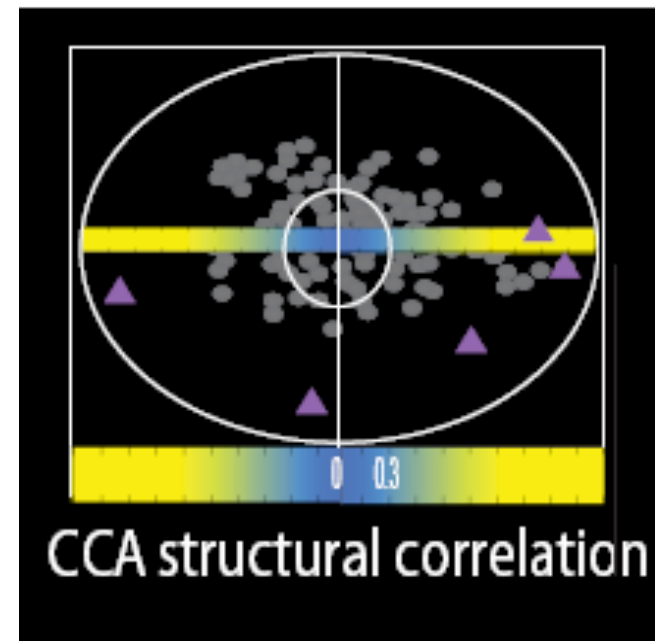


Strength of Pathway co-variation with environment



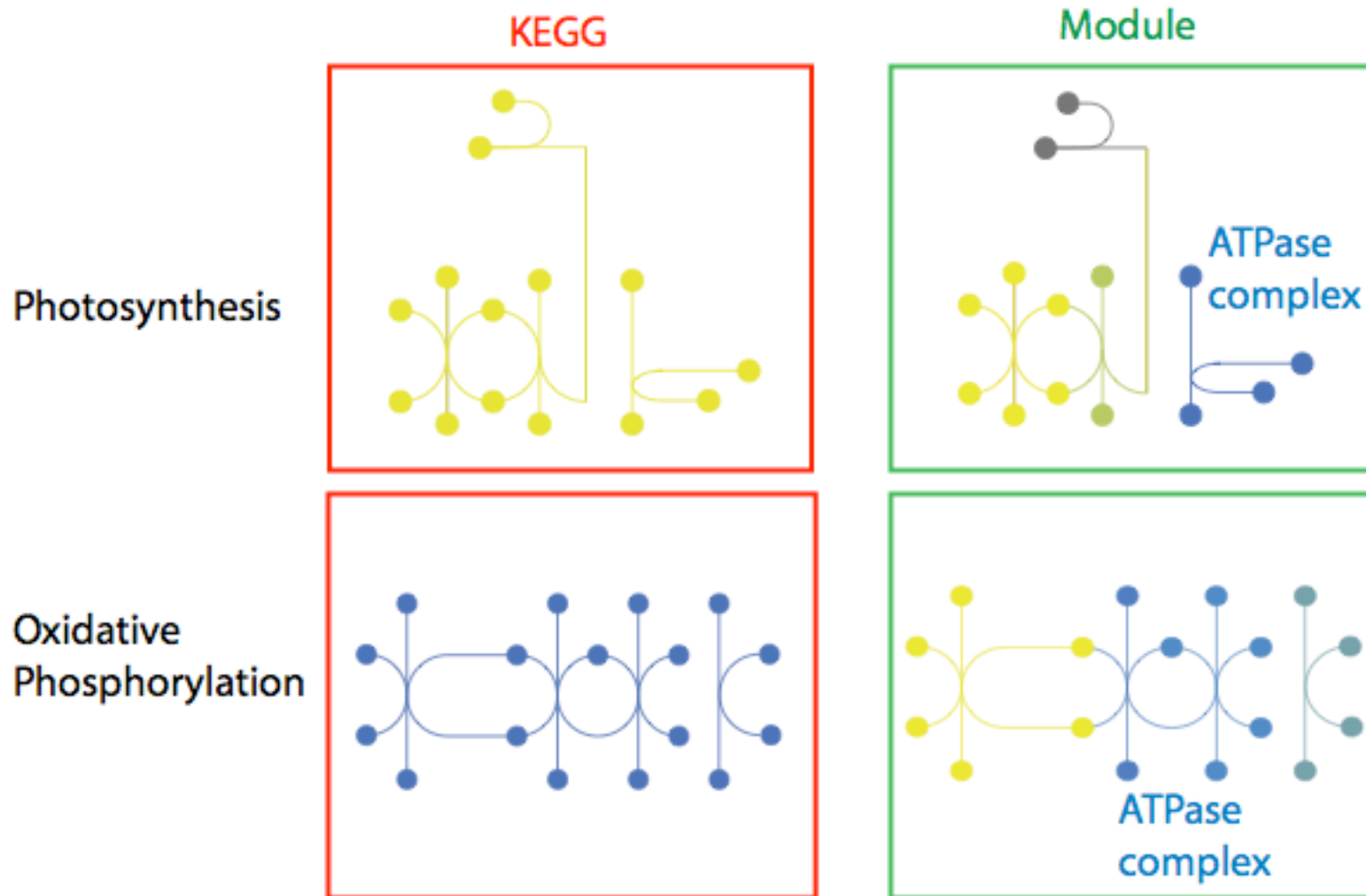
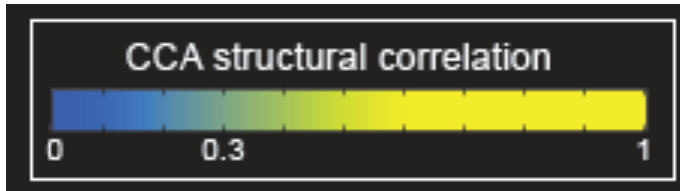
Environmentally
invariant

Environmentally
variant

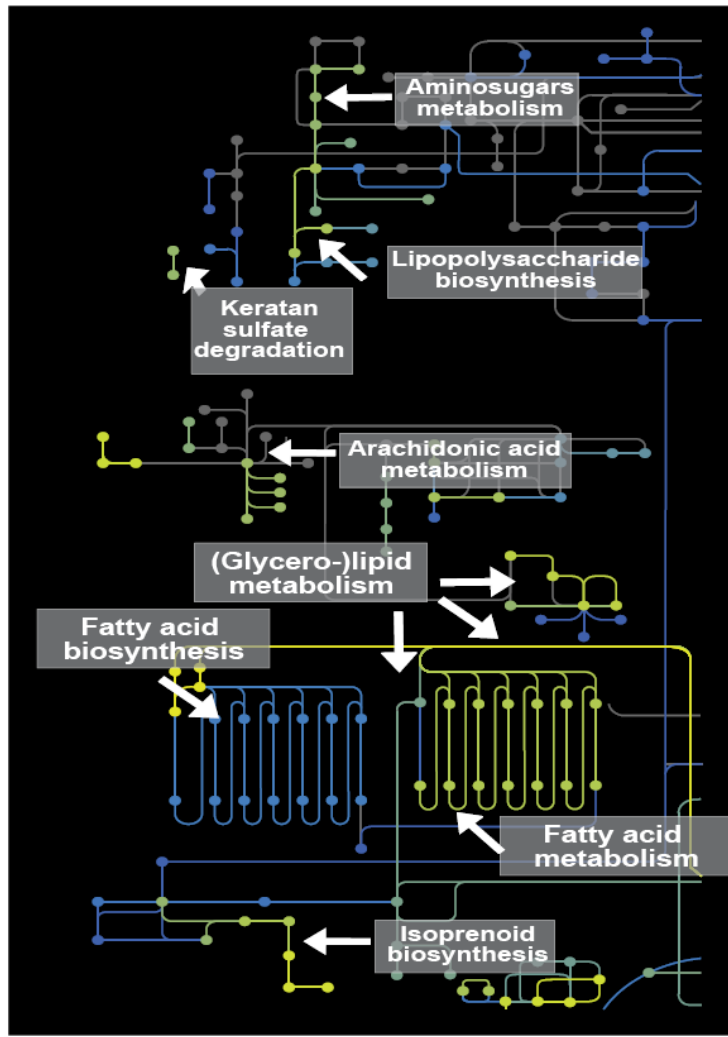


[Gianoulis et al., PNAS (in press, 2009)]

Conclusion #1: energy conversion strategy, temp and depth

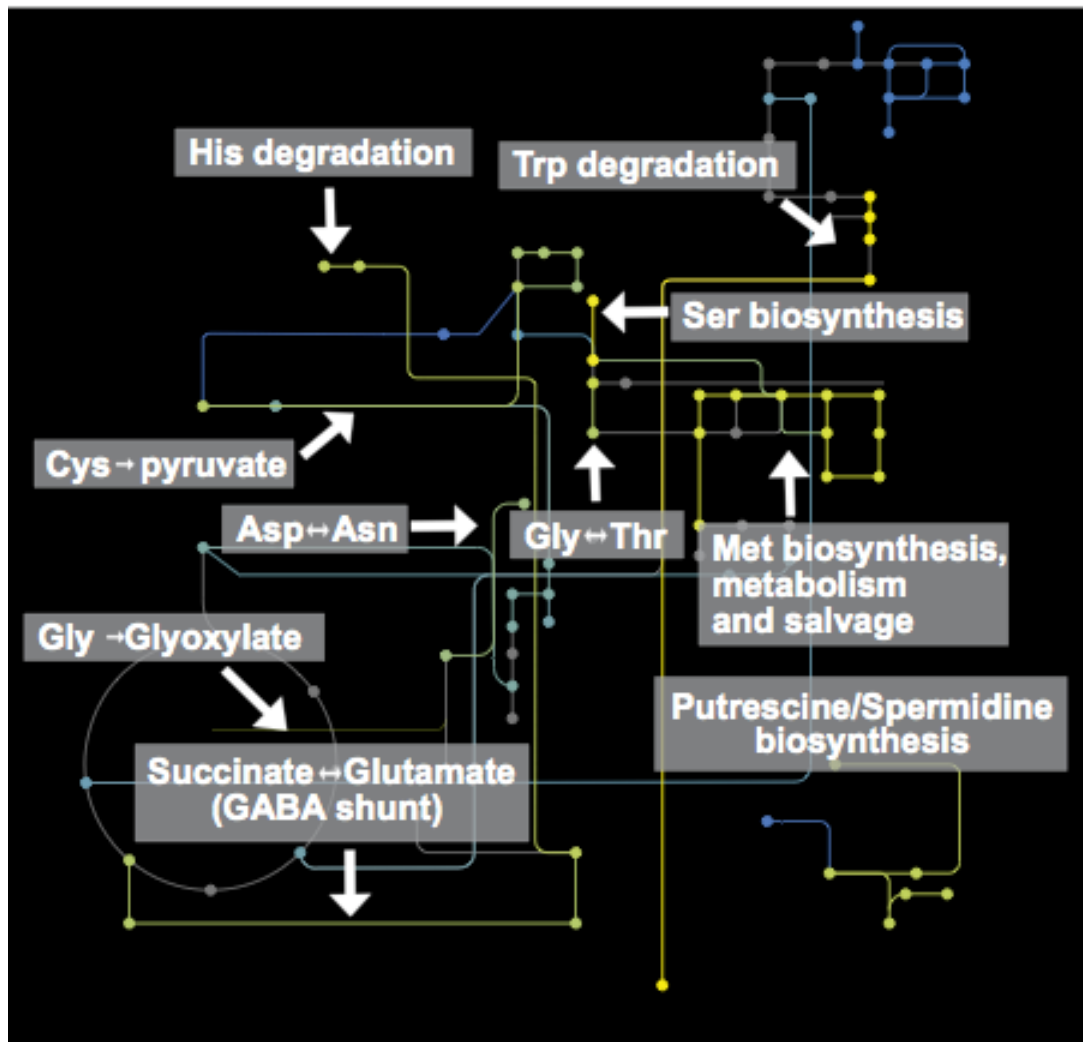


Conclusion #2: Outer Membrane components vary the environment



[Gianoulis et al., PNAS (in press, 2009)]

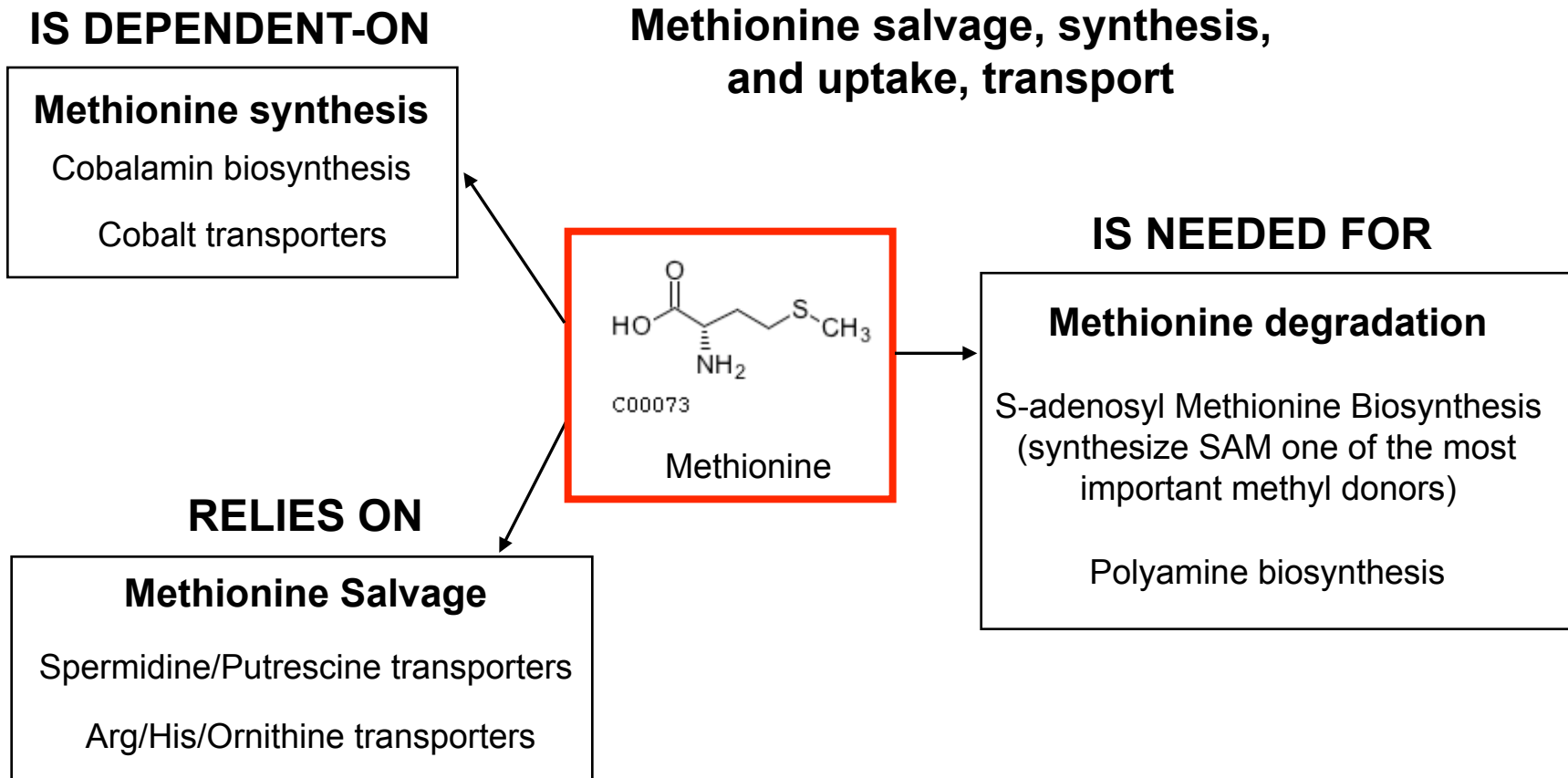
Conclusion #3: Covariation of AA biosynthesis and Import



Why is their fluctuation in amino acid metabolism? Is there a feature(s) that underlies those that are environmentally-variant as opposed to those which are not?

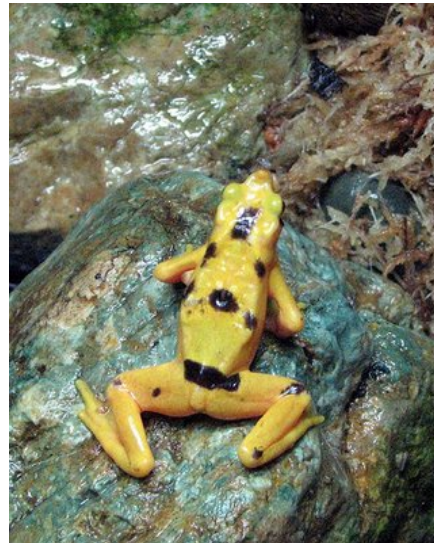
[Gianoulis et al., PNAS (in press, 2009)]

Conclusion #4: Cofactor (Metal) Optimization



[Gianoulis et al., PNAS (in press, 2009)]

Biosensors: Beyond Canaries in a Coal Mine



[Gianoulis et al., PNAS (in press, 2009)]

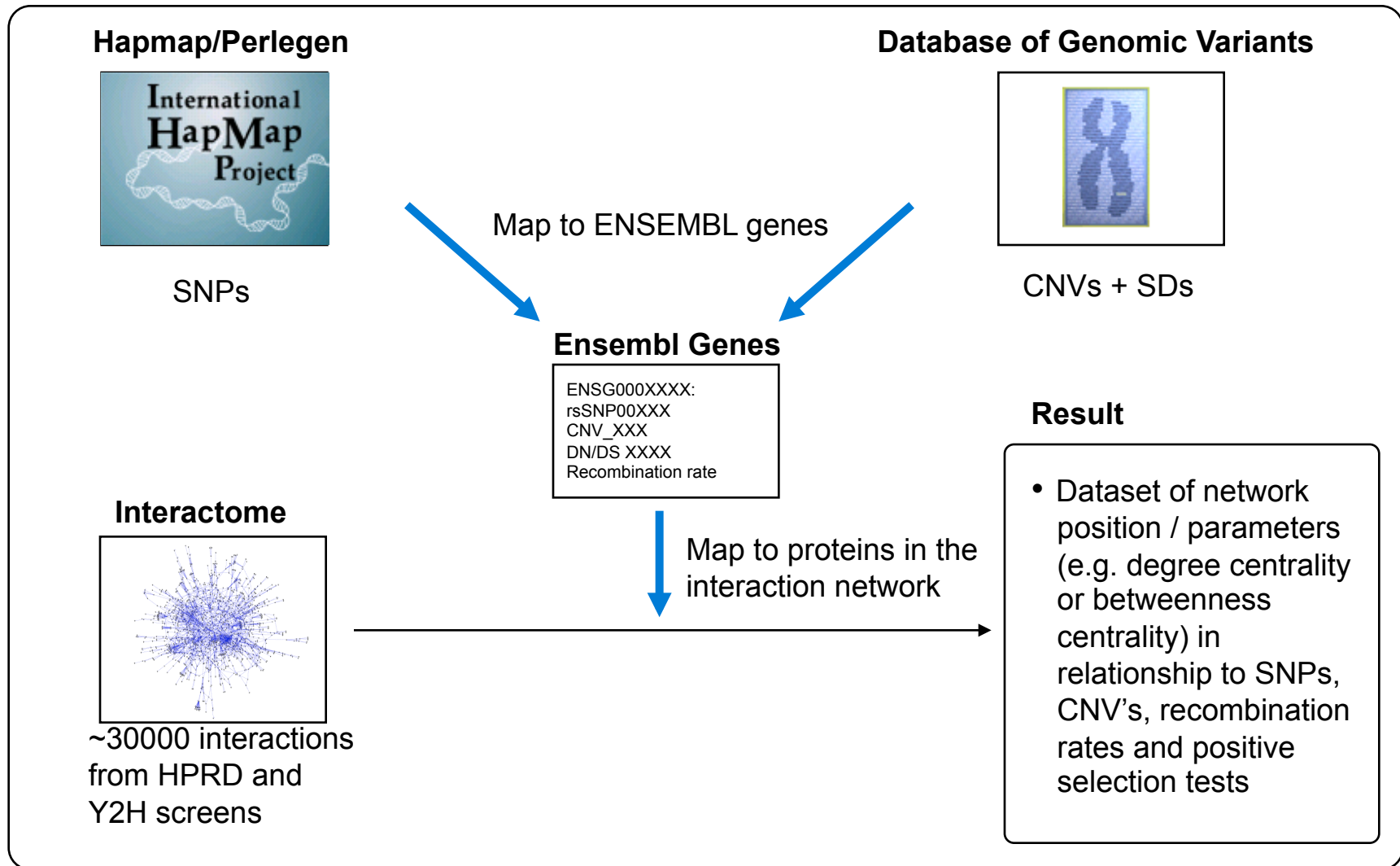
Networks & Variation

Which parts of the network vary most in sequence?
Which are under selection, either positive or negative?



METHODOLOGY: MAP SNP AND CNV DATA ONTO ENSEMBL GENES, AND THEN MAP ENSEMBL GENES TO THE KNOWN INTERACTOME

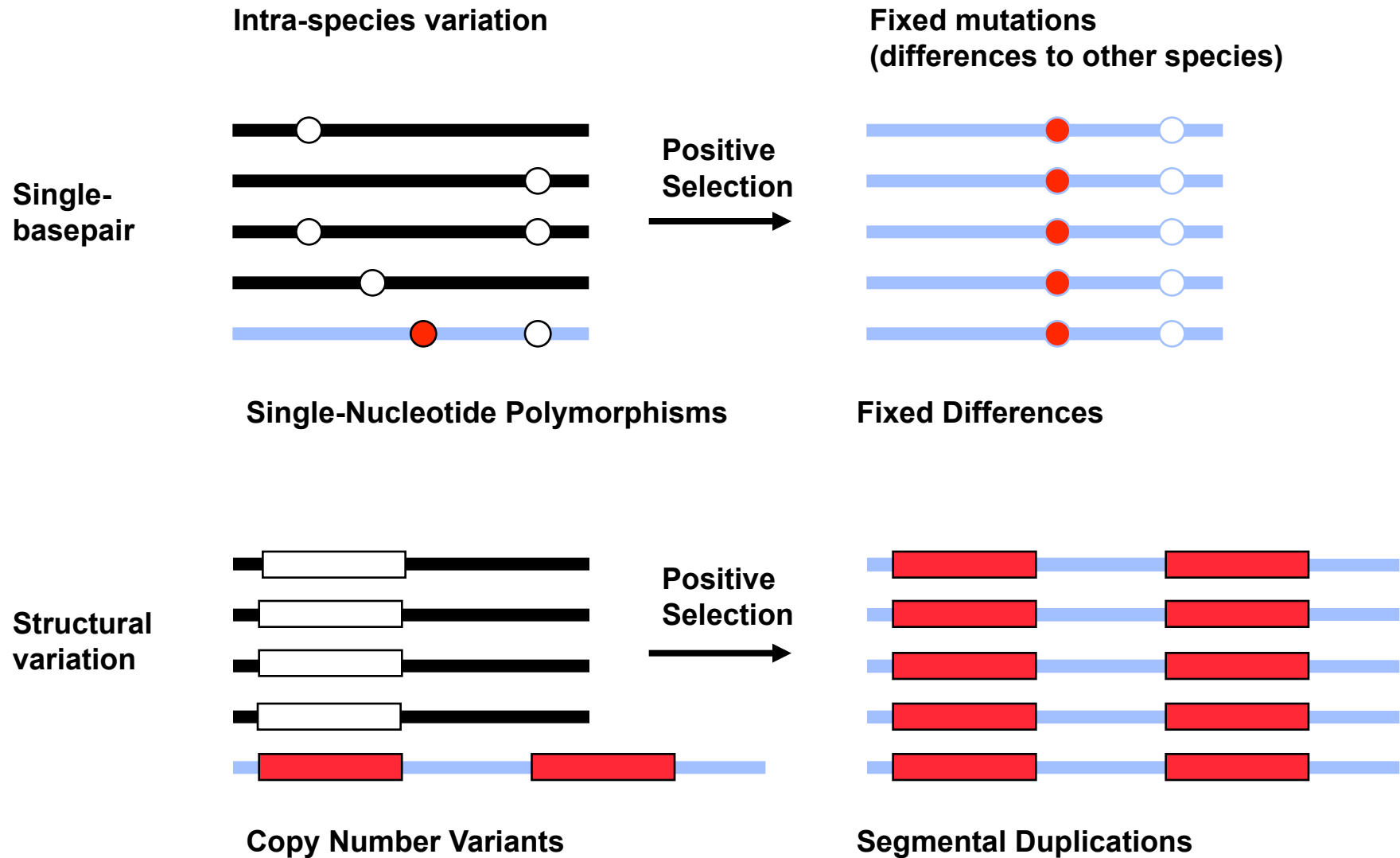
ILLUSTRATIVE



*From Nielsen et al. *PLoS Biol.* (2005) and Bustamante et al. *Nature* (2005)

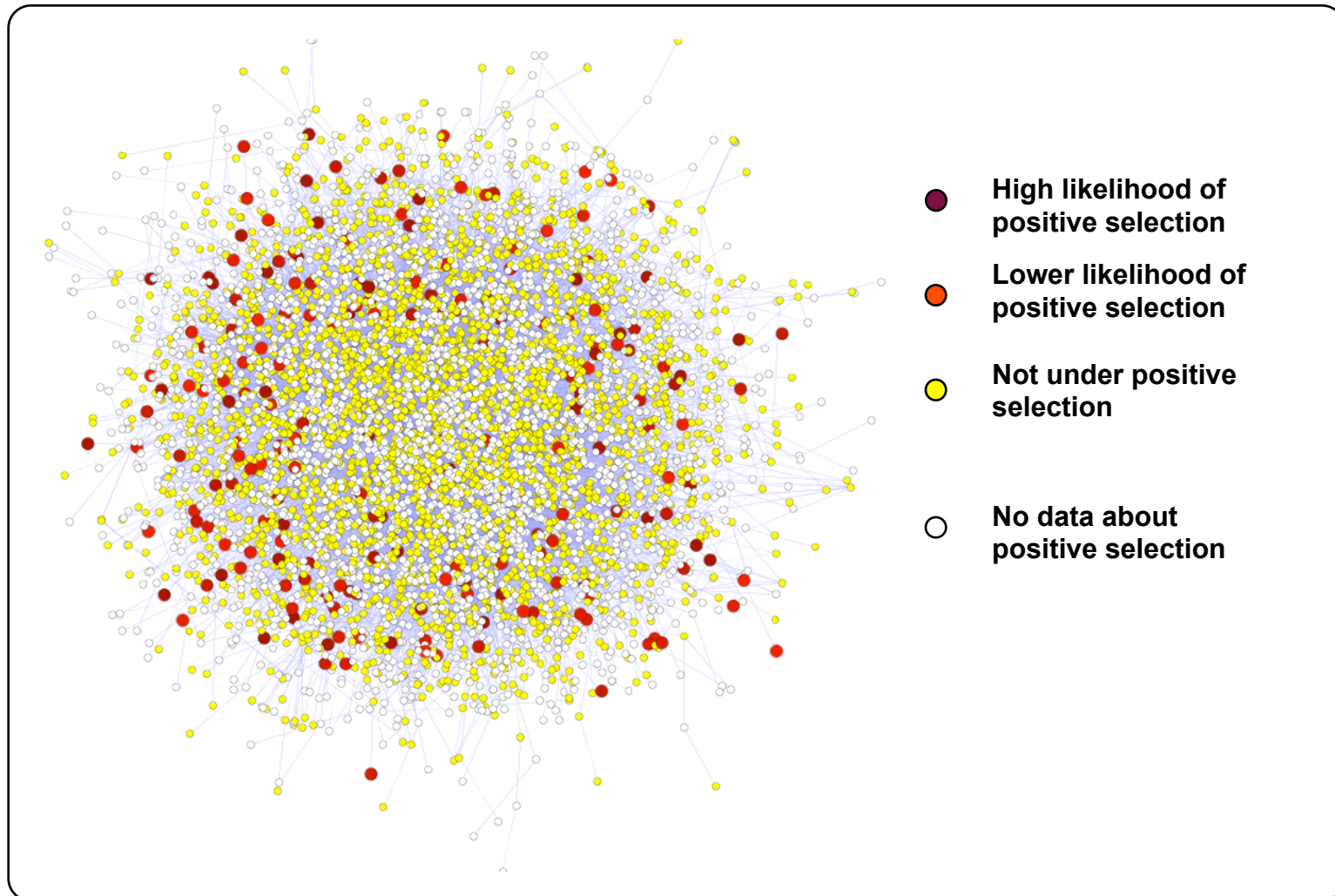
Source: PMK

ADAPTIVE EVOLUTION CAN BE SEEN ON TWO DIFFERENT LEVELS



POSITIVE SELECTION LARGELY TAKES PLACE AT THE NETWORK PERIPHERY

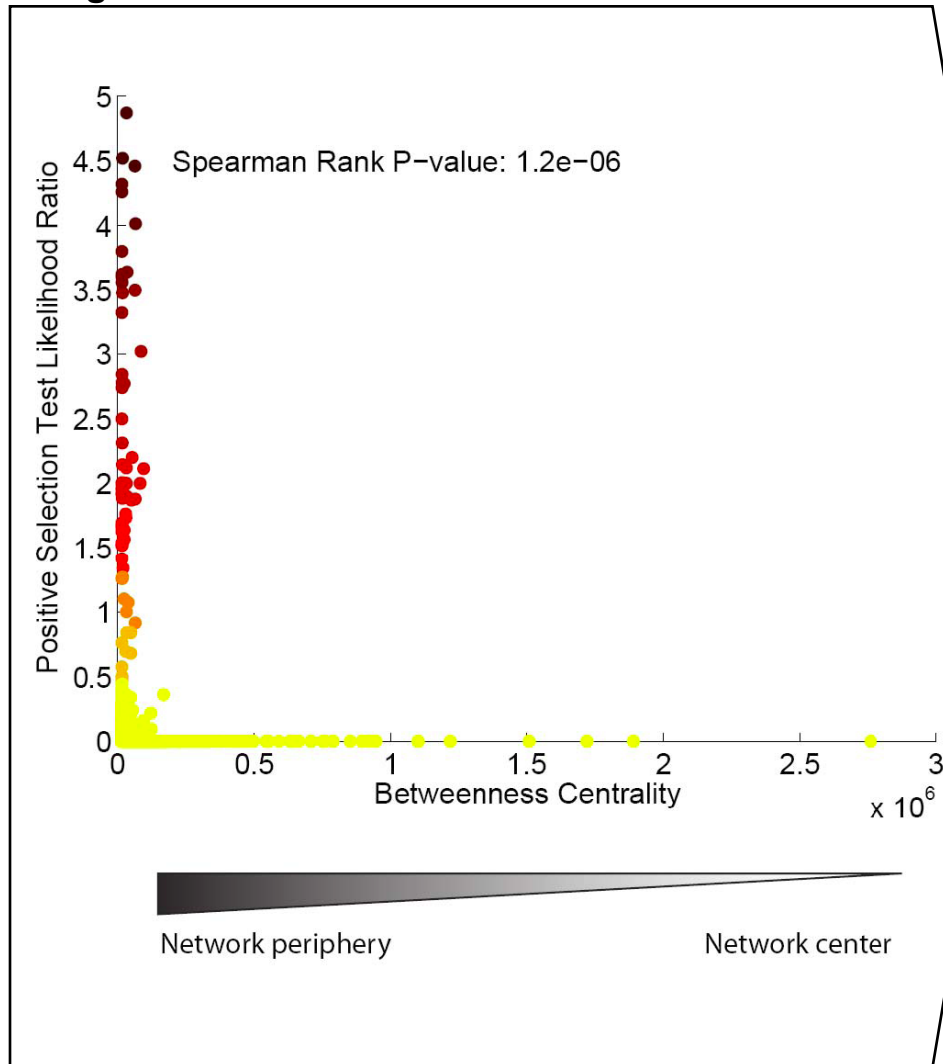
Positive selection in the human interactome



CENTRAL PROTEINS ARE LESS LIKELY TO BE UNDER POSITIVE SELECTION

▢ Hubs

Degree vs. Positive Selection



Reasoning

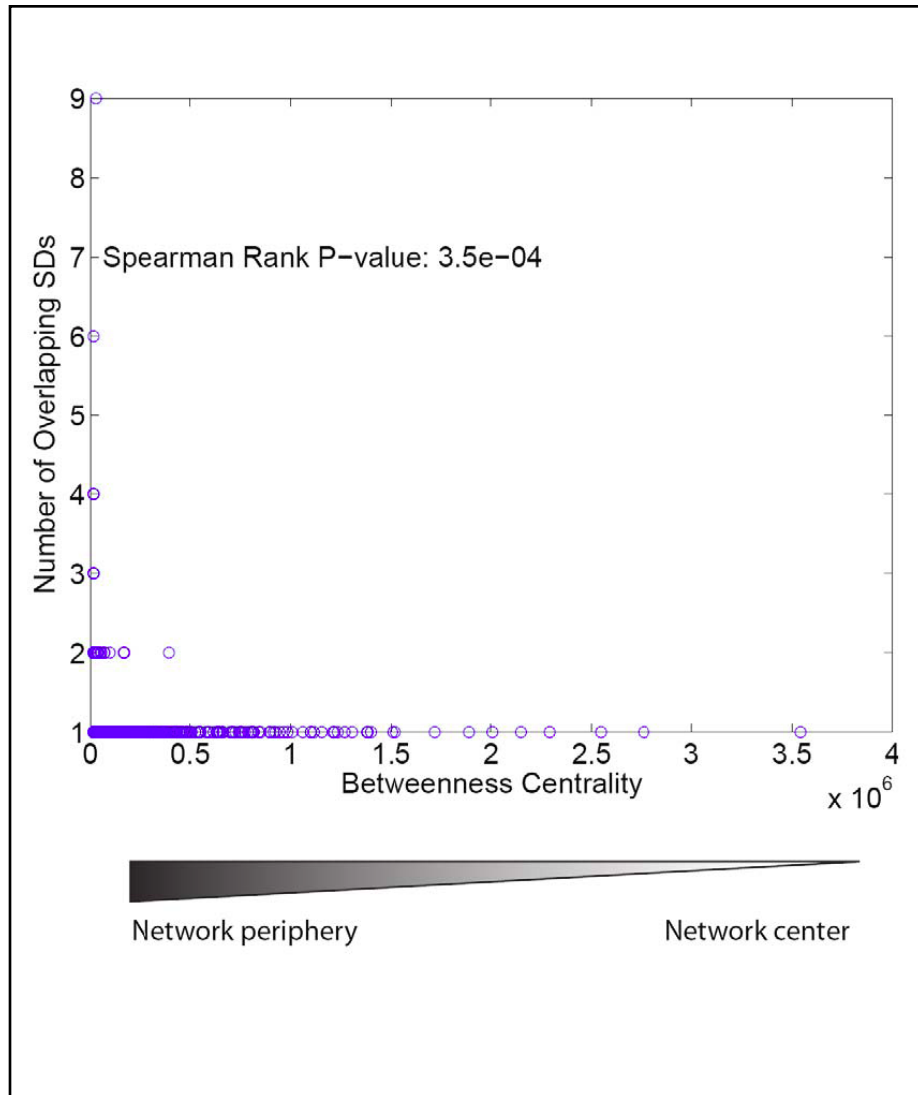
- Peripheral genes are likely to under positive selection, whereas hubs aren't
- This is likely due to the following reasons:
 - Hubs have stronger structural constraints, the network periphery doesn't
 - Most recently evolved functions (e.g. “environmental interaction genes” such as sensory perception genes etc.) would probably lie in the network periphery
- Effect is independent of any bias due to gene expression differences

*With a probability of over 80% to be positively selected as determined by Ka/Ks. Other tests of positive selection (McDonald Kreitmann and LDD) corroborate this result.

Source: Nielsen et al. *PLoS Biol.* (2005), Bustamante et al. *Nature* (2005), HPRD, Rual et al. *Nature* (2005), and Kim et al. *PNAS* (2007)

CENTRAL NODES ARE LESS LIKELY TO LIE INSIDE OF SDs

Centrality vs. SD occurrence



Reasoning

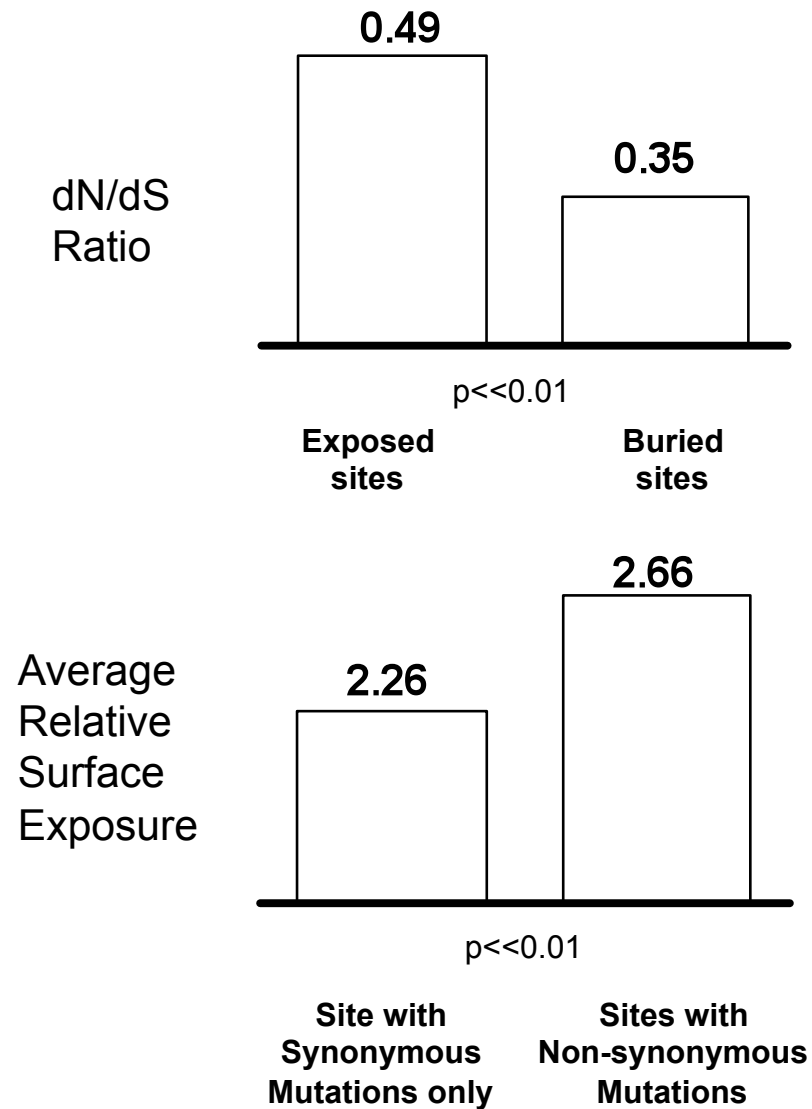
- This result also confirms our initial hypothesis – peripheral nodes tend to lie in regions rich in SDs.
- Since segmental duplications are a different mechanism of ongoing evolution, the less constrained peripheral proteins are enriched in them.
- Note that despite the small size of our dataset for known SD's we get significant correlations. It is to be expected that the correlations will get clearer as more data emerges*

*Specifically, a number of the SDs are likely not fixed, but rather common CNVs in the reference genome

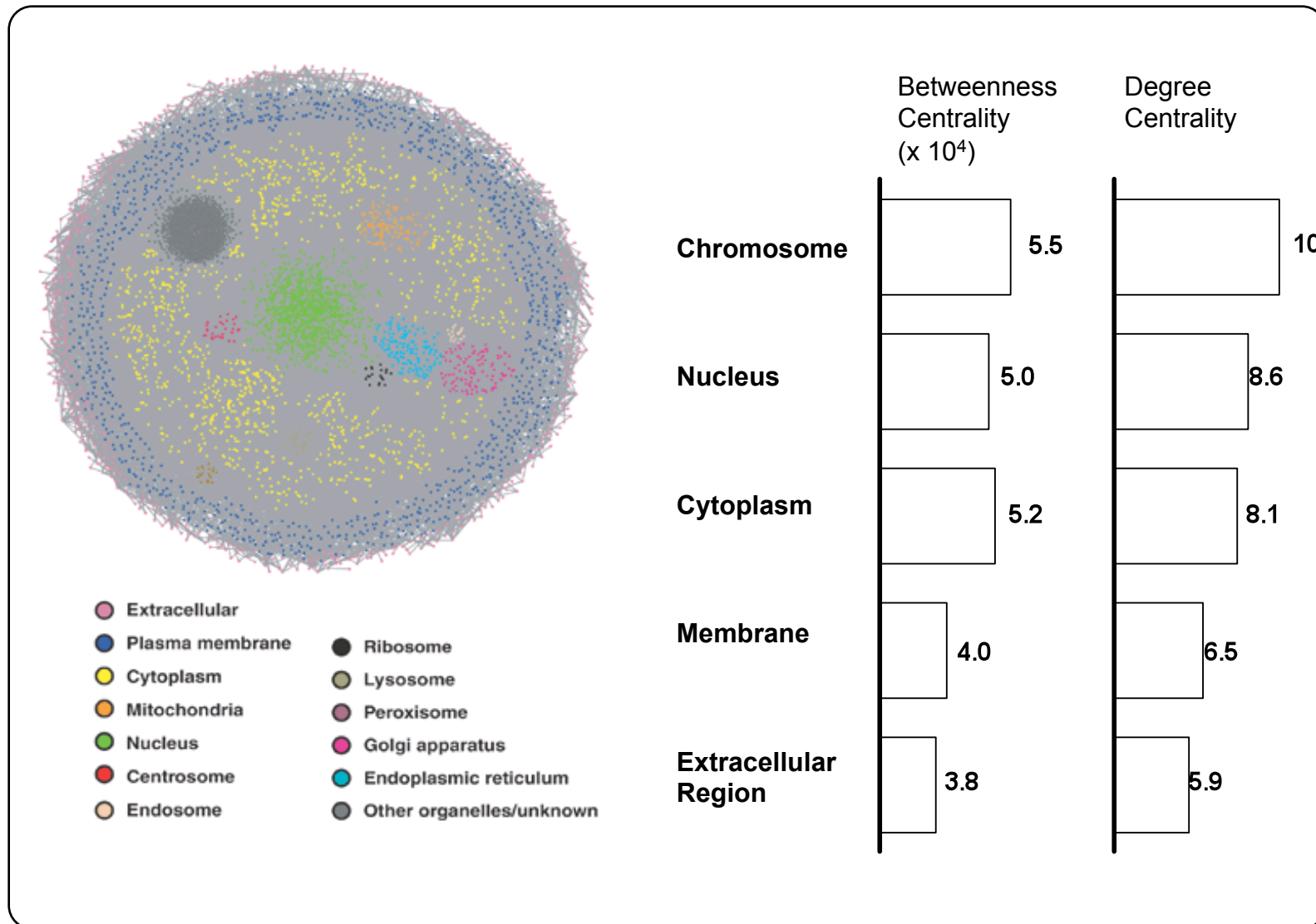
Source: Database of genetic variation, HPRD, Rual et al. *Nature* (2005), and Kim et al. *PNAS* (2007)

Why do we observe this? Perhaps central hub proteins are involved in more interactions & have more surface buried.

**BURIED SITES ARE
CONSERVED AND
MUCH LESS LIKELY
TO HARBOR NON-
SYNONYMOUS
MUTATIONS**



Another explanation: THE NETWORK PERIPHERY CORRESPONDS TO THE CELLULAR PERIPHERY



Source: Gandhi et al. (*Nature Genetics* 2006), Kim et al. PNAS (2007)

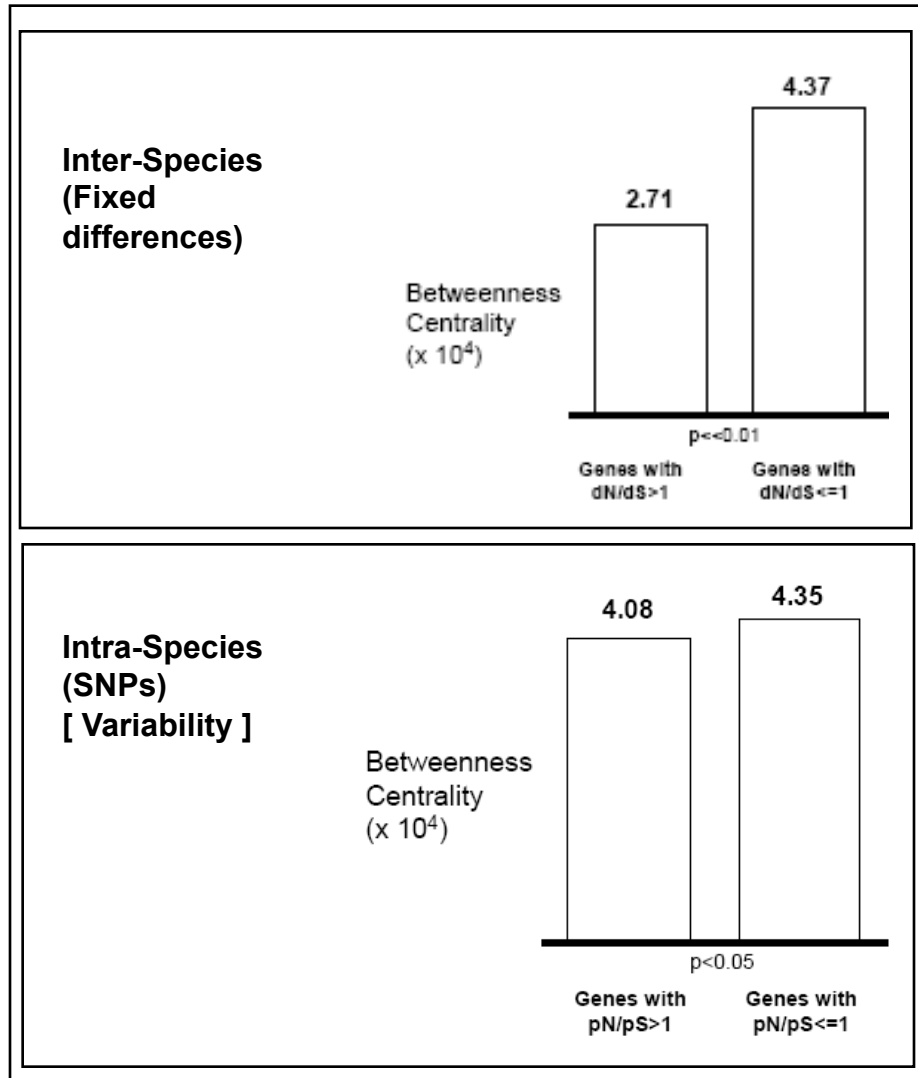
IS RELAXED CONSTRAINT OR ADAPTIVE EVOLUTION THE REASON FOR THE PREVALENCE OF BOTH SELECTED GENES AND SDs AT THE NETWORK PERIPHERY?

ILLUSTRATIVE

	Relaxed Constraint	Adaptive Evolution
Inter-Species Variation (Fixed differences)	<ul style="list-style-type: none">• Increases inter-species variation – more variable loci are under less negative selection• Can be seen in higher Ka/Ks ratio or SD occurrence	<ul style="list-style-type: none">• Increases inter-species variation – more variable loci are under less negative selection• Can be seen in higher Ka/Ks ratio or SD occurrence
Intra-Species Variation (Polymorphisms)	<ul style="list-style-type: none">• Increases intra-species variation – for the very same reason• Can be seen in both SNPs or CNVs	<ul style="list-style-type: none">• Should not have effects on intra-species variation

SOME, BUT NOT ALL OF THE SINGLE-BASEPAIR SELECTION AT THE PERIPHERY IS DUE TO RELAXED CONSTRAINT

Inter vs. Intra-Species Variation in Networks



Reasoning

- There is a difference in **variability** (in terms of SNPs) between the network periphery and the center
- However, this difference is much smaller than the difference in **selection**
- This most likely means, that part of the effect we're seeing is due to relaxed constraint (and higher variability)
- But, not the entire effect*

*But it's hard to quantify

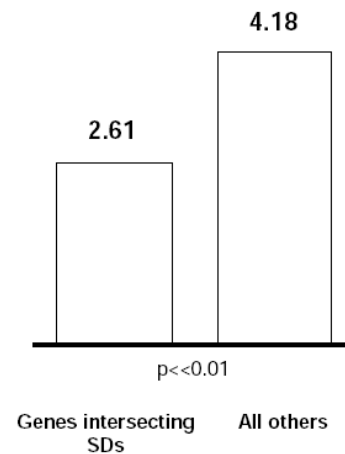
Source: Kim et al. (2007) PNAS

Similar Results for Large-scale Genomic Changes (CNVs and SDs)

Inter vs. Intra-Species Variation in Networks

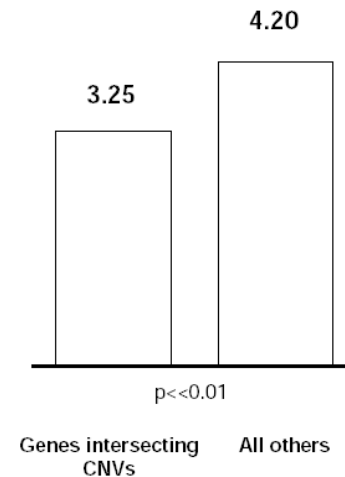
Inter-Species (SDs)

Betweenness Centrality ($\times 10^4$)



Intra-Species (CNVs) [Variability]

Betweenness Centrality ($\times 10^4$)



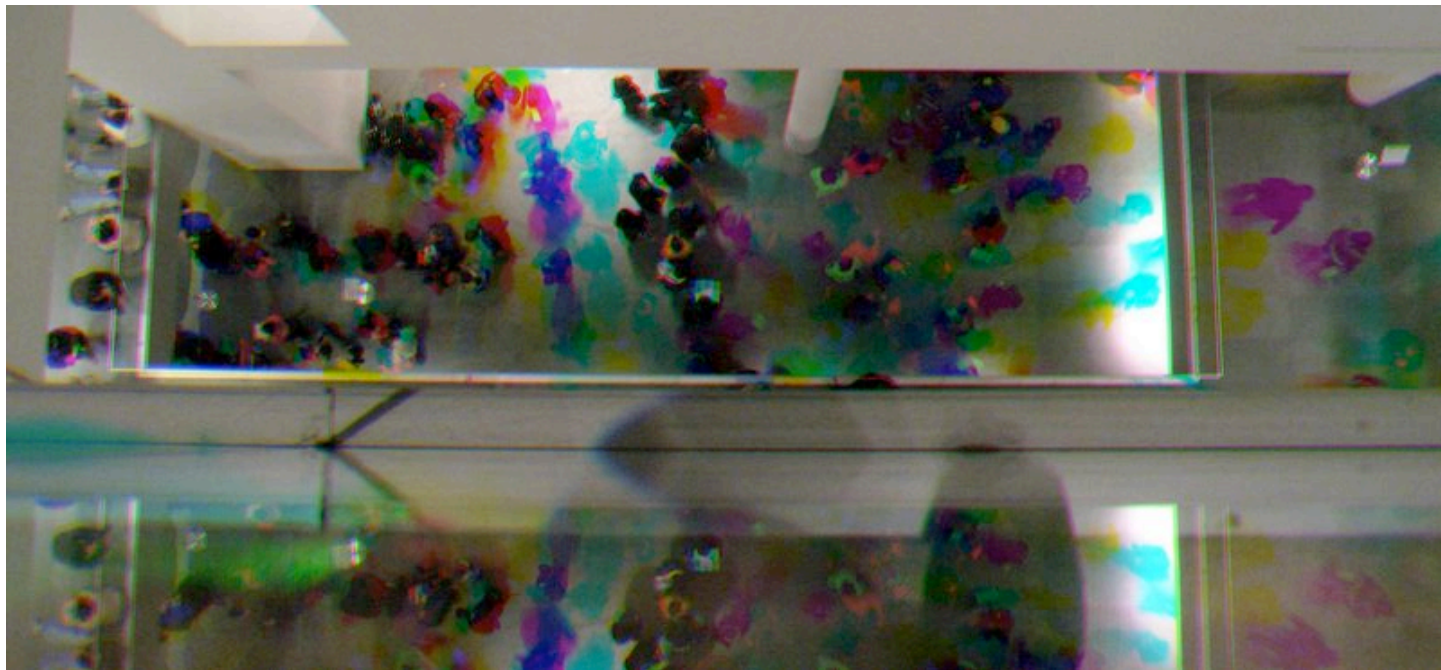
Reasoning

- There a small difference in **variability** (in terms of CNVs) between the network periphery and the center
- But, there is a (as shown before) marked difference in fixed (and hence, presumably, **selected**) SDs at the network periphery and center

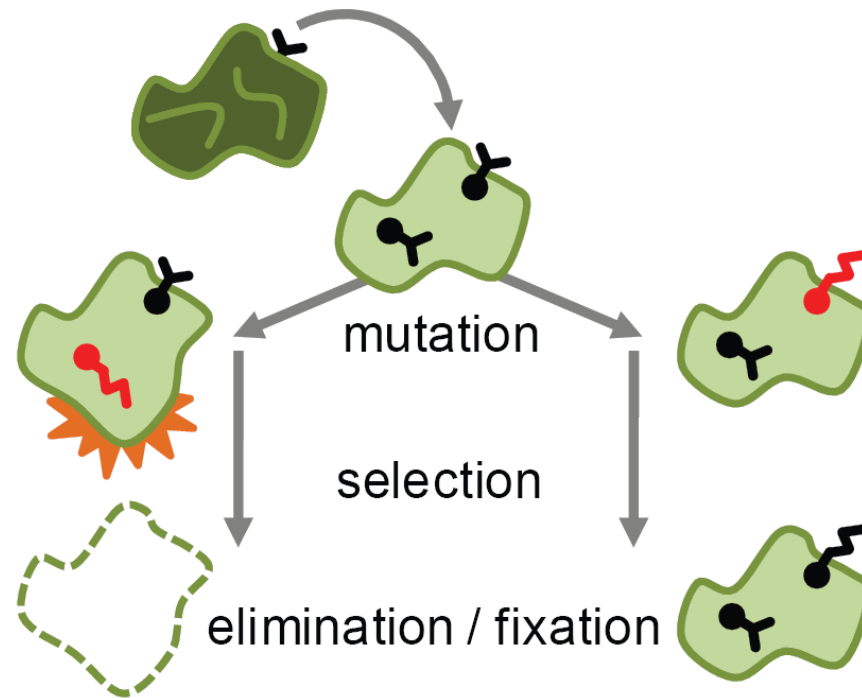
Networks & Variation 2

Which parts of the network vary most in sequence?

**More generally which features are
most correlated with evolutionary rate**



Protein evolutionary rate



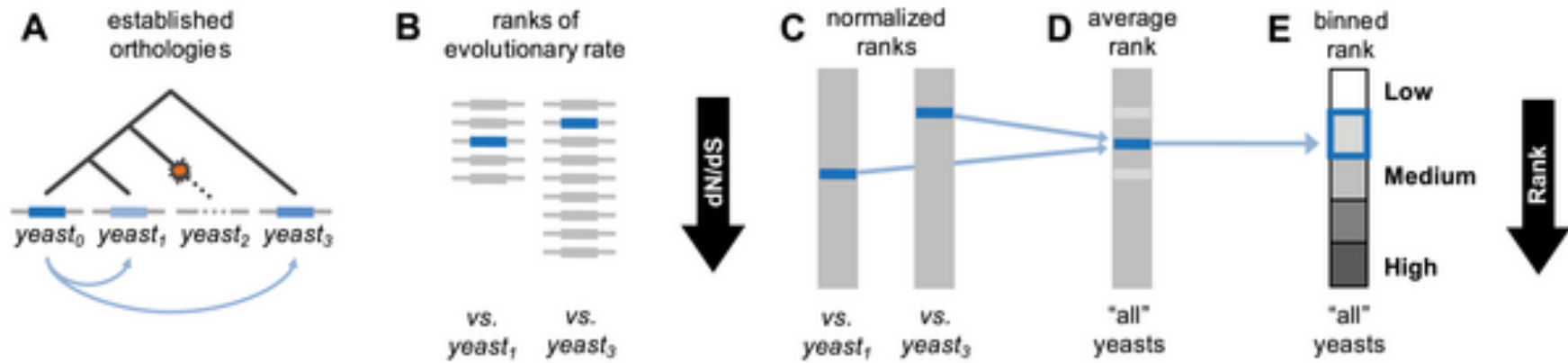
Evolutionary rate at the whole protein level

Why do some proteins evolve slowly, while others evolve quickly?

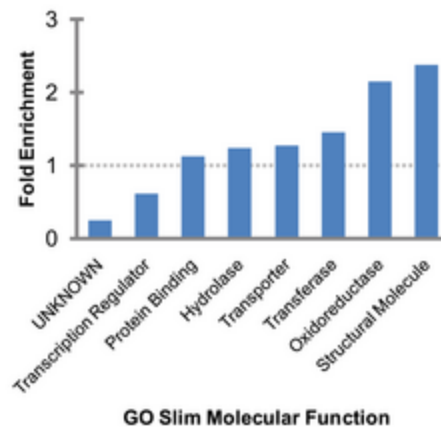
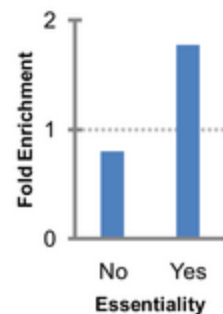
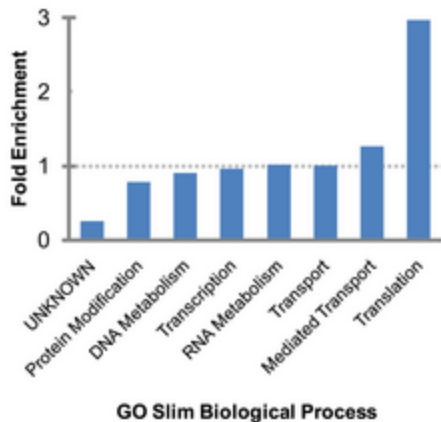
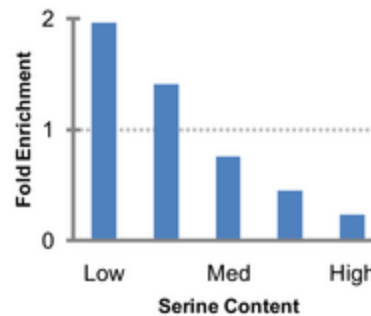
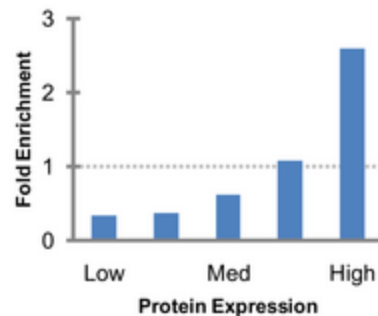
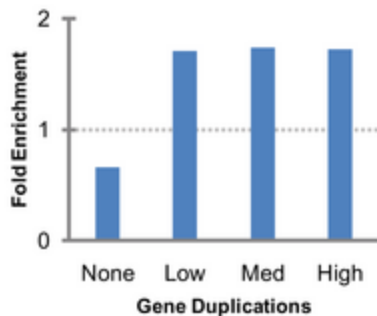
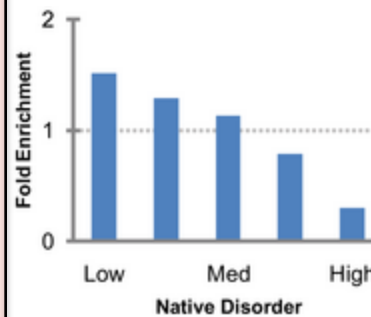
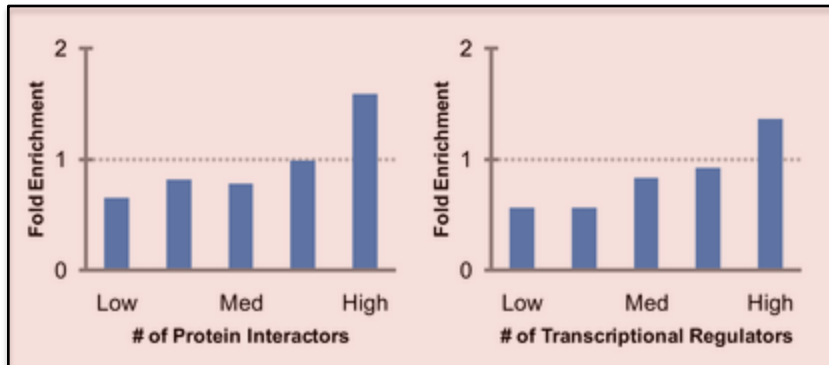
Many Gene Features Potentially Correlated with Evolutionary Rate

Meta-features	Features
Amino Acid Composition	Amino Acid Content (20 total attributes)
Structure (Physicochemical Properties)	Predicted helix content
	Predicted sheet content
	Predicted coil content
	Predicted native disorder
	Predicted transmembrane helix content
	Charge (pI)
	Hydrophobicity (Kyte-Doolittle)
Function	Aromaticity
	Size
	Biological process (GO slim)
	Molecular function (GO slim)
Abundance	Cellular compartment (GO slim)
	Absolute mRNA expression
	Protein expression
	Codon Adaptation Index (CAI)
Phenotype	Codon bias
	Essentiality
	Marginal essentiality
Network	Number of interactors
	Number of transcriptional regulators
Genome	Degree of gene duplication
	GC content

Assess Relationship of Many Different Features to Evolutionary Rate in Yeast



Simple Relationships between Individual Features & Evolutionary Rate

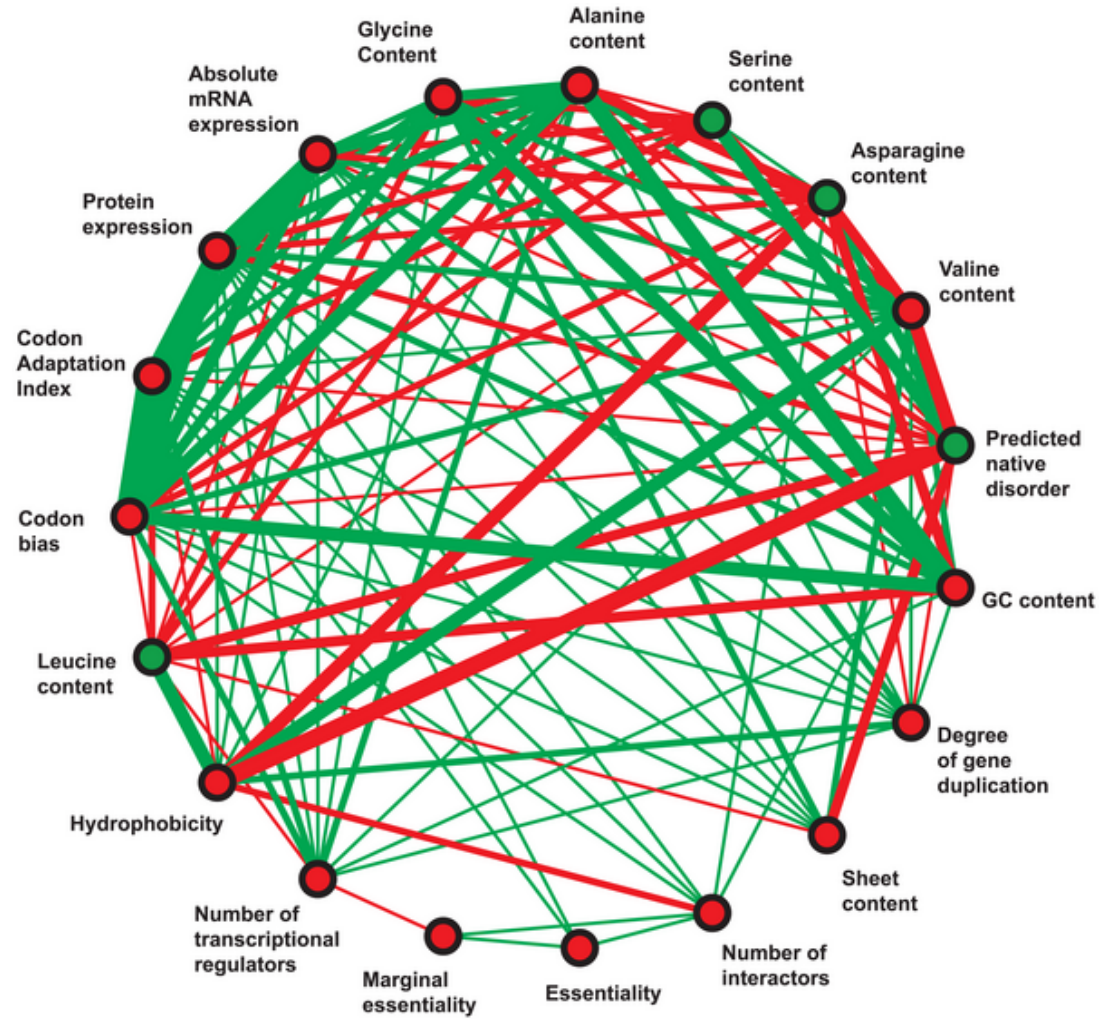


Feature Description	Rank Correlation with Evolutionary Rate
Codon bias	-0.578
Codon adaptation index	-0.557
Protein expression	-0.486
Absolute mRNA expression	-0.467
Gly content	-0.401
Ala content	-0.390
Ser content	0.366
Asn content	0.317
Val content	-0.293
Native disorder	0.251
GC content	-0.242
Degree of gene duplication	-0.206
Sheet content	-0.191
Number of interactors	-0.160
Essentiality	-0.147
Marginal essentiality	-0.146
# of transcriptional regulators	-0.142
Hydrophobicity	-0.141
Leu content	0.105
Gln content	0.081

doi:10.1371/journal.pcbi.1000413.t002

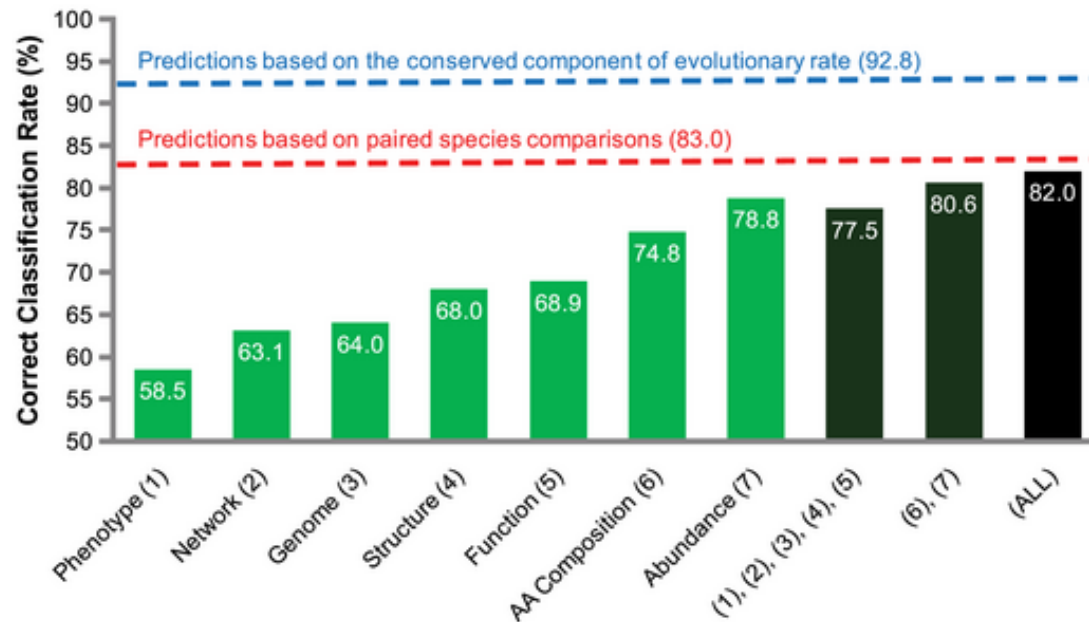
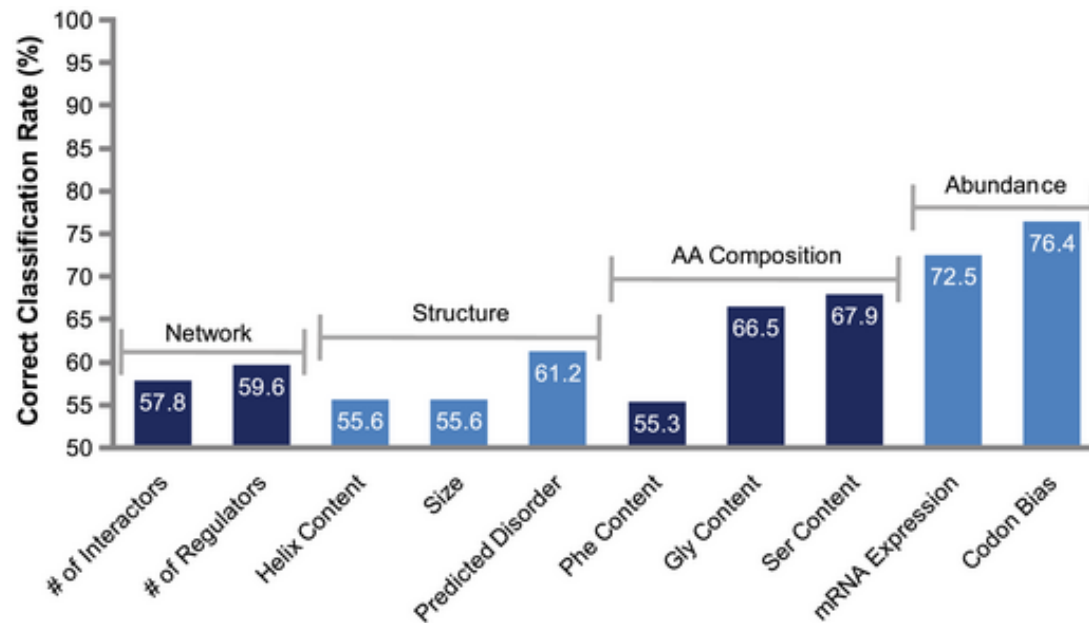
[Xia et al. ('09). Plos CB]

Correlations among top correlates of evolutionary rate



[Xia et al. ('09). Plos CB]

Information contribution of protein features & “meta-features” in the task of predicting slowly evolving proteins



[Xia et al. ('09). Plos CB]

Outline: Molecular Networks

- Why Networks?
- Predicting Networks (yeast)
 - ◇ Propagating known information
- Dynamics & Variation of Networks
 - ◇ Across cellular states (yeast)
 - ◇ Across environments (in prokaryotes)
- Protein Networks & Variation (yeast & humans)



Conclusions on Networks: Predictions



- Predicting Networks
 - ◇ Extrapolating from the Training Set
 - ◇ Principled ways of using known information in the fullest possible fashion
 - Prediction Propagation
 - Multi-level learning

Conclusions: Network Dynamics across Cellular States



- Merge expression data with Networks
- Active network markedly different in different conditions
- Identify transient hubs associated with particular conditions
- Use these to annotate genes of unknown function

Conclusions: Networks Dynamics across Environments



- Developed and adapted techniques to connect quantitative features of environment to metabolism.
- Applied to available aquatic datasets, we identified footprints that were predictive of their environment (potentially could be used as biosensor).
- Strong correlation exists between a community's energy conversion strategies and its environmental parameters (e.g. temperature and chlorophyll).
- Suggest that limiting amounts of cofactor can (partially) explain increased import of amino acids in nutrient-limited conditions.

Conclusions: Connecting Networks & Variation



- We find ongoing evolution (positive selection) at the network periphery.
 - ◇ This trend is present on two levels:
 - On a sequence level, it can be seen as positive selection of peripheral nodes
 - On a structural level, it can be seen as the pattern of SDs that display significantly higher allele frequencies in non-central genes
 - ◇ 2 possible mechanisms for this : adaptive evolution at cellular periphery & relaxation of structural constraints at the network periphery
 - We show that the latter can only explain part of the increased variability

•

Conclusions: Connecting Networks & Variation 2



- Evolutionary rate is related to network positioning
- However, only a weak relationship, with more association with abundance and composition



- an automated web tool

tYNA

(vers. 2 :

"TopNet-like

Yale Network Analyzer")

tYNA

Getting started API WSDL Download tYNA Installation guide Plugins for Cytoscape Contact Known problems

You are logged in as kevin. [Logout](#) View: Simple Advanced

List Owned Biological networks with (Attribute name) = (Attribute value) List

Workspace manager

Load an existing network

Load: 14. Uetz 2000 yeast two ...

Into: workspace 0

Categorized by: NII

Load

Current working networks in your workspaces:

Workspace 0: statFilter(degrees, geq, 1, value, neighbors=false, intersection("Uetz 2000 yeast two hybrid", "Ito 2001 yeast two hybrid"))

Workspace 1: (empty)

Workspace 2: (empty)

Workspace 3: (empty)

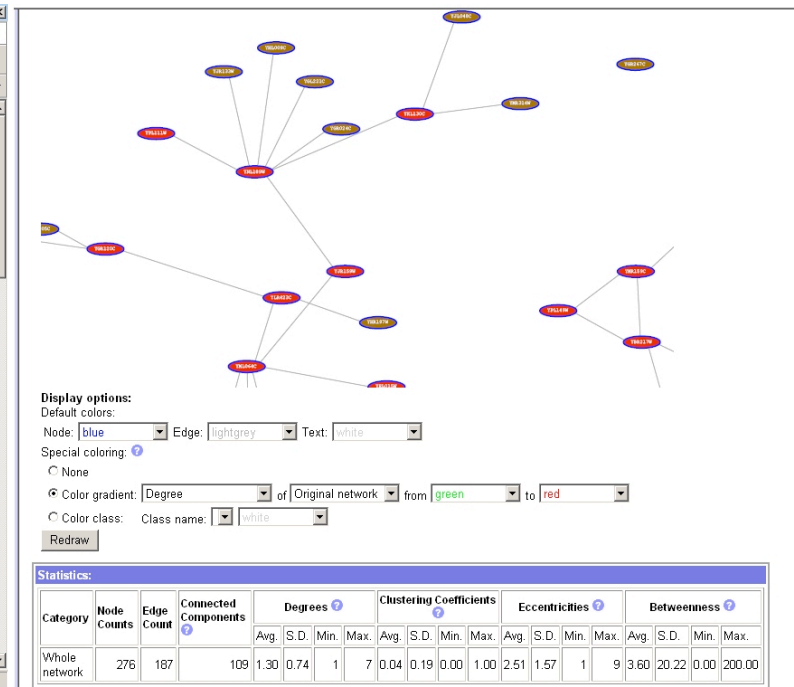
Multiple network analysis

Networks in database (upload download)

ID	Name	Creator	Creation date	
14	Uetz 2000 yeast two hybrid	kevin	21-Feb-06	Delete
15	Ito 2001 yeast two hybrid	kevin	21-Feb-06	Delete
16	Ho 2002 pull down	kevin	21-Feb-06	Delete
17	Gavin 2002 pull down	kevin	21-Feb-06	Delete
18	Jansen 2003 PIT	kevin	21-Feb-06	Delete
19	MIPS yeast PPI	kevin	21-Feb-06	Delete
21	BIND yeast data	kevin	21-Feb-06	Delete
22	DIP yeast data	kevin	21-Feb-06	Delete
23	Kim 2006 structural interaction	kevin	21-Feb-06	Delete
24	Han 2004 FYI data	kevin	21-Feb-06	Delete
25	Luscombe 2004 regulatory	kevin	21-Feb-06	Delete

Categories in database (upload download)

ID	Name	Creator	Creation date
----	------	---------	---------------

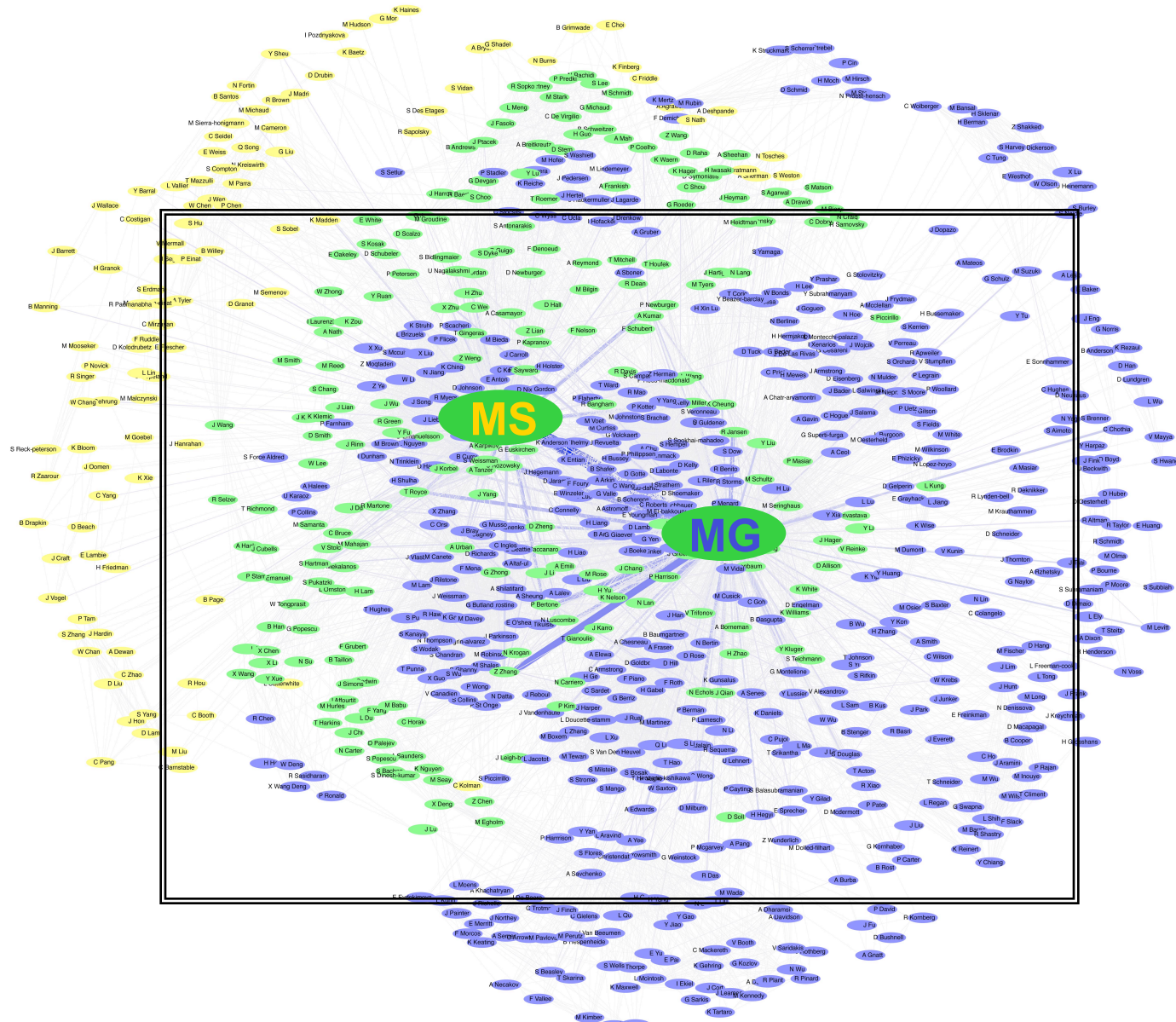


Normal website + Downloaded code (JAVA)
+ Web service (SOAP) with Cytoscape plugin

[Yu et al., NAR (2004); Yip et al. Bioinfo. (2006);
Similar tools include Cytoscape.org, Idekar, Sander et al]

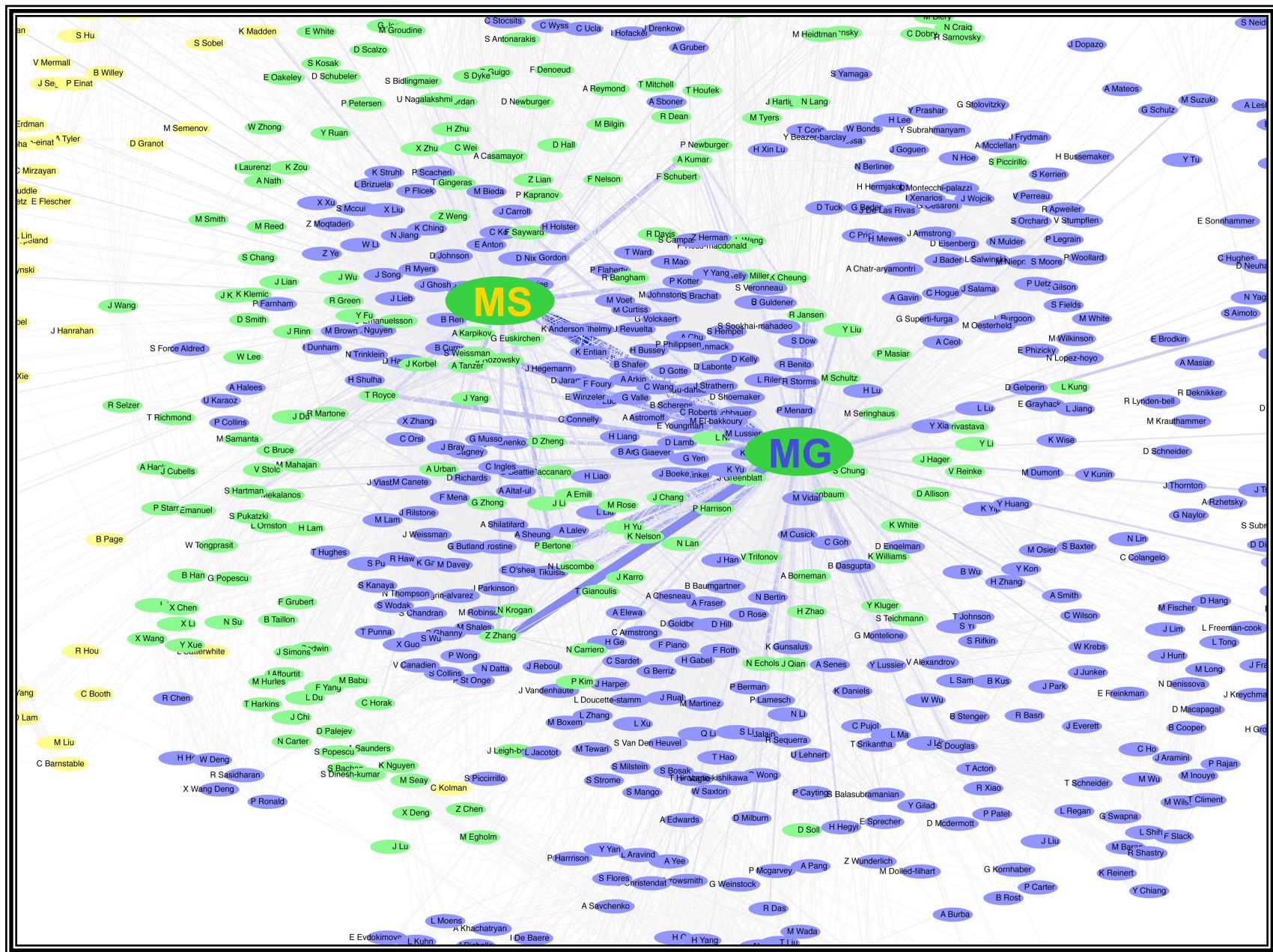
Acknowledgements

Networks.GersteinLab.org



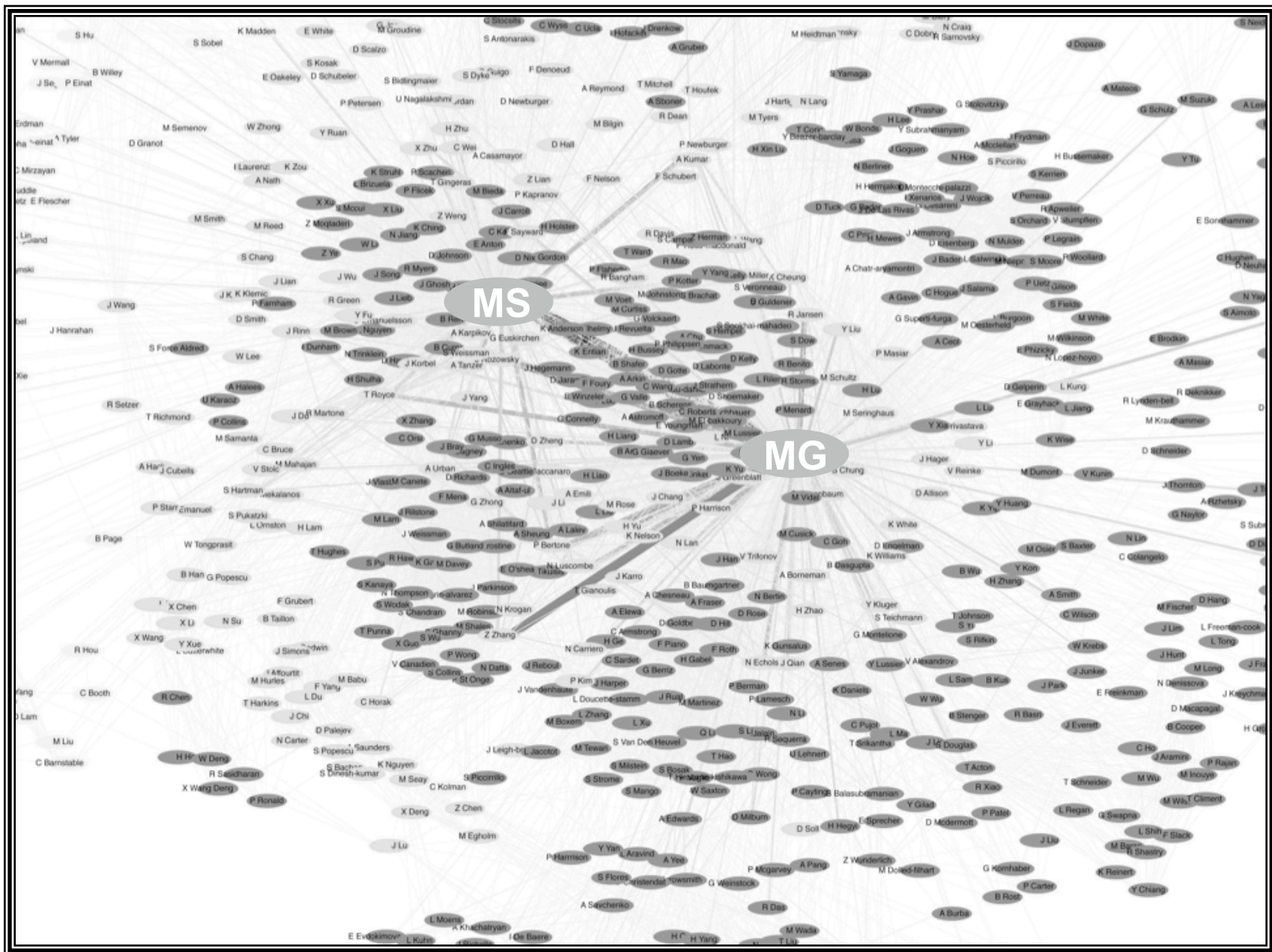
Acknowledgements

Networks.GersteinLab.org



Acknowledgements

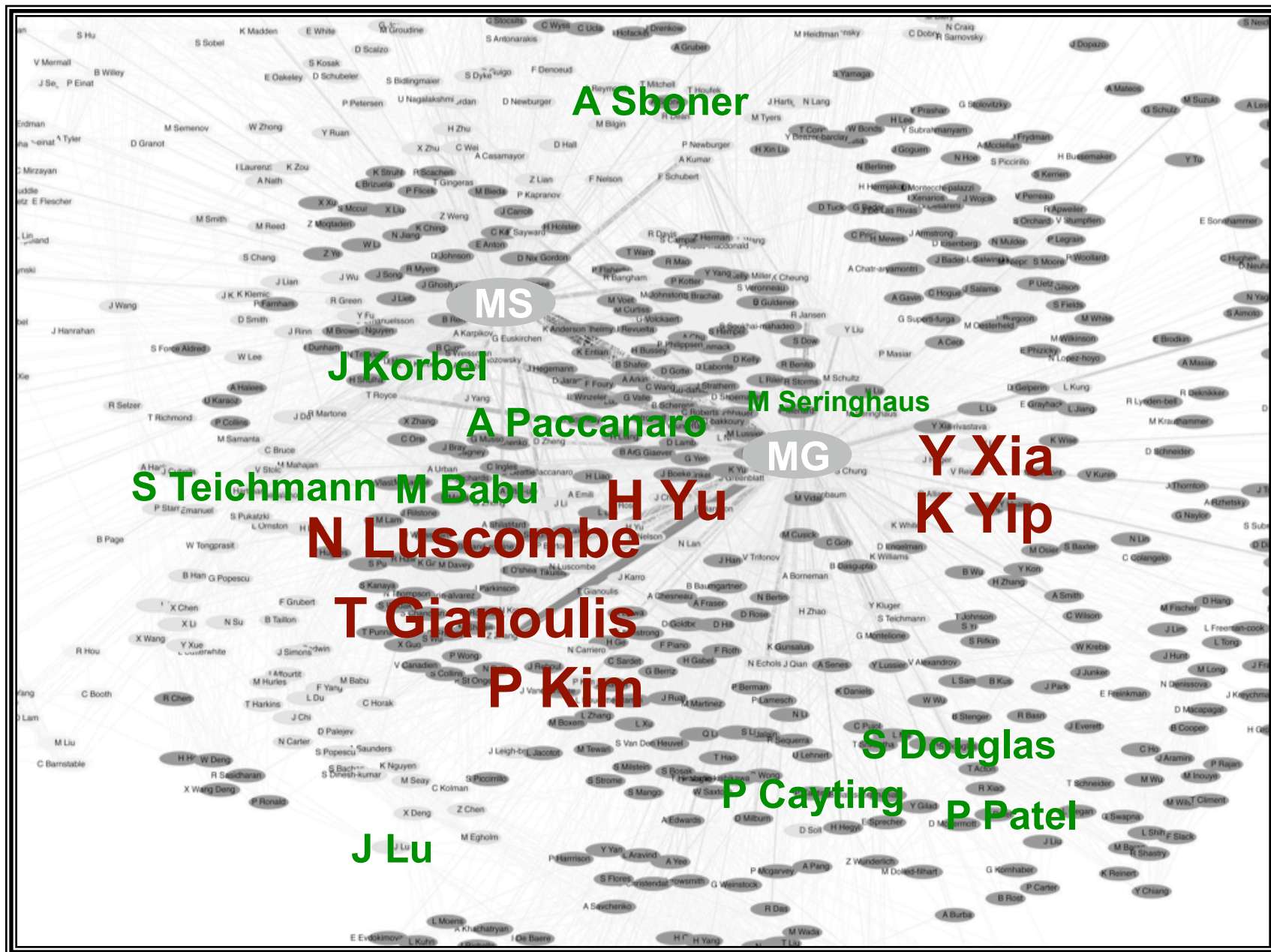
Networks.GersteinLab.org



P Bork, J Raes,
E Franzosa

Acknowledgements
Networks.GersteinLab.org

Job opportunities currently
for postdocs & students



More Information on this Talk

TITLE: Understanding Protein Function on a Genome-scale through the Analysis of Molecular Networks

SUBJECT: Networks

DESCRIPTION:

Summit on Systems Biology 2009, The Microbial World and Beyond, Richmond, VA, 2009.05.19, 13:00–14:00; [I:**3RDSUMMIT**] (Long networks talk, adding in for the first time: **evolrate***. Fits easily into 50' w. 10' questions. PPT works on mac & PC and has many photos.)

(Paper references in the talk were mostly from Papers.GersteinLab.org. The above topic list can be easily cross-referenced against this website. Each topic abbrev. which is starred is actually a papers "ID" on the site. For instance, the topic **pubnet*** can be looked up at <http://papers.gersteinlab.org/papers/pubnet>)

PERMISSIONS: This Presentation is copyright Mark Gerstein, Yale University, 2008. Please read permissions statement at <http://www.gersteinlab.org/misc/permissions.html> . Feel free to use images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).

PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see <http://streams.gerstein.info> . In particular, many of the images have particular EXIF tags, such as **kwpotppt** , that can be easily queried from flickr, viz: <http://www.flickr.com/photos/mbgmbg/tags/kwpotppt> .